JCTC Journal of Chemical Theory and Computation

Computation of Absolute Hydration and Binding Free Energy with Free Energy Perturbation Distributed Replica-Exchange Molecular Dynamics

Wei Jiang,[†] Milan Hodoscek,[‡] and Benoît Roux*,[†],[§]

Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Building 221 Argonne, Illinois 60439, Center for Molecular Modeling, National Institute of Chemistry, Hajdrihova 19, SI-1000 Ljubljana, Slovenia, and Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, University of Chicago, 929 57th Street, Chicago, Illinois 60637

Abstract: Distributed Replica (REPDSTR) is a powerful parallelization technique enabling simulations of a group of replicas in a parallel/parallel fashion, where each replica is distributed to different nodes of a large cluster [*Theor. Chem. Acc.* **2003**, *109*, 140]. Here, we use the framework provided by REPDSTR to combine a staged free energy perturbation protocol with replica-exchange molecular dynamics (FEP/REMD). The structure of REPDSTR, which allows multiple parallel input/output (I/O), facilitates the treatment of replica-exchange to couple the N window simulations corresponding to different values of the thermodynamic coupling parameters. As a result, each of the N synchronous window simulations benefit from the sampling carried out by the N-1 others. As illustrative examples of the FEP/REMD strategy, calculations of the absolute hydration and binding free energy of small molecules were performed using the biomolecular simulation program CHARMM adapted for the IBM Blue Gene/P platform. The computations show that a FEP/REMD strategy significantly improves the sampling and accelerates the convergence of absolute free energy computations.

Introduction

The design of accurate and rapid methods for calculating the free energy of solvation and binding of small molecules is one of the central goals of computer simulations. Progress in this area could, for example, help speed up the development and prediction of new therapeutic molecules and accelerate drug discovery. In principle, this can be achieved using free energy perturbation (FEP) calculations based on molecular dynamics (MD) simulations of atomic models including explicitly the solvent molecules. To enable robust and effective computations of absolute solvation free energies and of binding free energies, a staged FEP/MD simulation protocol based on a step-by-step decomposition of the total reversible work was developed.[1−4] The staged protocol effectively breaks down the complete free energy calculation into several independent MD simulations, which are easy to distribute over independent compute nodes. Nevertheless, a treated based on independent simulations does not exploit all the available information in the computations for the sampling. Furthermore, such a treatment does not make the best possible use of supercomputing platforms allowing massively distributed processes.

One possible strategy to improve the FEP/MD is to couple the staged free energy simulations using replica-exchange methodologies.[5−12] With REMD, the rate at which the configurations of a system are being explored can be considerably enhanced by attempting coordinate swapping between independent simulations generated at different temperatures[5,6,11,12] or with different Hamiltonians.[7−12] Of particular interest, replica-exchange has been combined in Monte Carlo simulations to compute relative free energies.[7,13] This suggests that REMD might be a promising avenue to further improve FEP/MD simulations. The Distributed Replica (REPDSTR) technique recently developed and implemented in the biomolecular program CHARMM[14] by Hodoscek and co-workers[15,16] allows multiple MD simulations on similar systems to be simultaneously performed in a highly efficient parallel/parallel mode. In the REPDSTR implementation, each independent MD simulation is treated as an independent replica that occupies a group of processors, and each has its "private" input/output (I/O), including the input script. This is illustrated in Figure 1. REPDSTR is in contrast to the more standard parallel MD, where only the root process deals with the I/O. For this reason, the parallel scaling in REPDSTR is primarily determined by the number of CPUs occupied by each replica, rather than by the total number of CPUs in use. The multiple I/O and parallel structure of REPDSTR greatly facilitate the usage of replica-exchange protocols, with infrequent communication between the separate simulations. It follows that an implementation of a staged FEP protocol with replica-exchange between the different thermodynamic window simulations is relatively transparent and easy to manage. One advantage is that the resulting free energy

* Corresponding author e-mail: roux@uchicago.edu.
† Argonne National Laboratory.
‡ National Institute of Chemistry, Slovenia.
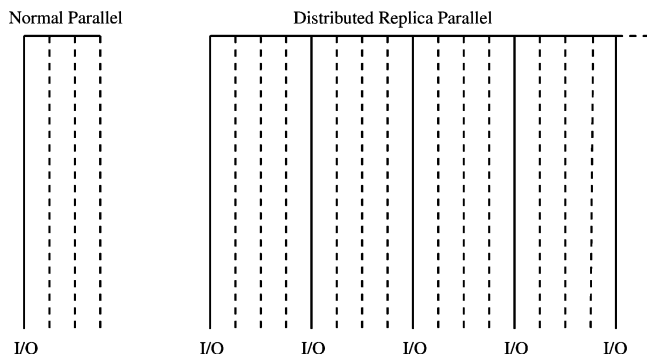§ University of Chicago.

Figure 1. Multiple I/O structure of distributed replica parallel.

perturbation distributed replica-exchange molecular dynamics (FEP/REMD) algorithm is controlled within a single program rather than via an external job-driven script. It is therefore well adapted for massively distributed supercomputing platforms.

In this communication, we describe and implement a computational scheme based on the REPDSTR of CHARMM for computing the absolute hydration and binding free energies within a staged FEP protocol using a replica-exchange MD simulations with λ-swap moves. It is shown that this FEP/REMD scheme improves the sampling efficiency and convergence of free energy computations. FEP/REMD is illustrated with the calculation of the absolute hydration free energy of water and benzene and the absolute binding free energy of camphor-P450 complex.

Computational Details

A. Staged FEP Simulation Protocol. The staged FEP/MD simulation protocol with biasing restraints was described in refs 1−4. Briefly, the insertion of the ligand into the binding pocket is done in three steps, with the help of the three thermodynamic coupling parameters, λ_{rep}, λ_{dis}, and λ_{elec}, controlling the nonbonded interaction of the molecule with its environment. One additional parameter, λ_{rstr}, is used to control the translational and orientational restraints. The potential energy is expressed in terms of the four coupling (window) parameters

$$U(\lambda_{rep}, \lambda_{dis}, \lambda_{elec}, \lambda_{rstr}) = U_0 + U_{rep}(\lambda_{rep}) + \lambda_{dis}U_{dis} + \lambda_{elec}U_{elec} + \lambda_{rstr}U_{rstr} \quad (1)$$

where U_0 is the potential of the system with the noninteracting ligand, U_{rep} and U_{dis} are the shifted Weeks-Chandler-Anderson (WCA) repulsive and dispersive components of the Lennard-Jones potential (introduced in ref 1), U_{elec} is the electrostatic contribution, and u_r denotes the restraining potential that helps improve phase space sampling. The repulsive contribution ΔG_{rep} corresponds to the process

$$U(\lambda_{rep} = 0, \lambda_{dis} = 0, \lambda_{elec} = 0, \lambda_{rstr} = 1) \rightarrow \\ U(\lambda_{rep} = 1, \lambda_{dis} = 0, \lambda_{elec} = 0, \lambda_{rstr} = 1) \quad (2)$$

the dispersive contribution ΔG_{dis} corresponds to the process

$$U(\lambda_{rep} = 1, \lambda_{dis} = 0, \lambda_{elec} = 0, \lambda_{rstr} = 1) \rightarrow \\ U(\lambda_{rep} = 1, \lambda_{dis} = 1, \lambda_{elec} = 0, \lambda_{rstr} = 1) \quad (3)$$

and the electrostatic contribution ΔG_{elec} corresponds to the process

$$U(\lambda_{rep} = 1, \lambda_{dis} = 0, \lambda_{elec} = 0, \lambda_{rstr} = 1) \rightarrow \\ U(\lambda_{rep} = 1, \lambda_{dis} = 1, \lambda_{elec} = 1, \lambda_{rstr} = 1) \quad (4)$$

The free energy ΔG_{rstr} corresponds to the process

$$U(\lambda_{rep} = 1, \lambda_{dis} = 1, \lambda_{elec} = 1, \lambda_{rstr} = 1) \rightarrow \\ U(\lambda_{rep} = 1, \lambda_{dis} = 1, \lambda_{elec} = 1, \lambda_{rstr} = 0) \quad (5)$$

The insertion of the ligand or solute into the bulk is calculated according to the same protocol but without the restraint.

B. REPDSTR Implementation. With the REPDSTR module of CHARMM,[15,16] each λ-staging FEP window is treated as a replica. The hydration free energy calculation is separated into three types of jobs, corresponding to the contributions from the repulsive, dispersive, and electrostatic nonbonded interactions. The binding free energy calculation is separated into four types of jobs, corresponding to the repulsive, dispersive, electrostatic, and restraint contributions. Each replica from the set of replicas within a given type occupies multiple processors, in a parallel/parallel mode. Figure 2 shows the REPDSTR implementation of each interaction type as well as the corresponding coupling parameters and replica labels.

The replica-exchange algorithm follows the conventional Metropolis MC exchange probability with λ-swap moves

$$P(\lambda_{type}^i \leftrightarrow \lambda_{type}^j) = \\ \min\{1, e^{-[U(\lambda_{type}^i, \mathbf{X}^i) + U(\lambda_{type}^j, \mathbf{X}^j) - U(\lambda_{type}^j, \mathbf{X}^i) - U(\lambda_{type}^i, \mathbf{X}^j)]/k_B T}\} \quad (6)$$

where U denotes the total potential energy of the underlying replica, and λ_{type}^i and λ_{type}^j denote the staging parameters. When a move is accepted, then the λ values are exchanged (λ-swap). During the REMD simulation, alternating pairs are considered for exchange attempts. Figure 2 shows these replica pairs for each type of interaction, with even and odd denoting the "alternating" replica exchange mode. This λ-swap replica-exchange protocol could easily be extended to the windows of umbrella sampling simulations.

C. MD Simulation. All the FEP/MD simulations were carried out on the IBM Blue Gene/P cluster Intrepid of Argonne National Laboratory using version c36a1 of the CHARMM program,[14] which was modified for the present study. For the hydration calculations each replica (window) occupies 16 CPUs, while for the binding free energy calculation each replica occupies 64 CPUs. To decrease the computational cost without compromising accuracy, reduced systems of the bulk solution and of the binding site were simulated in which the influence of the surrounding was incorporated by a mean-field treatment. The Spherical Solvent Boundary Potential (SSBP)[17] was used for the bulk simulation, and the Generalized Solvent Boundary Potential (GSBP) was used for the binding site simulations.[18] The hydration FEP calculations were done with 400 explicit water molecules and SSBP at 300 K. The systems were propagated with a 2 fs time step using Langevin dynamics. Calculations based on simulation of 40 and 100 ps with different replica-exchange frequencies were generated and compared. To estimate the statistical convergence of the calculations, 10 independent FEP/REMD simulations runs were performed consecutively starting from the configuration saved at the end of the previous run.

For the binding free energy calculations of the camphor-P450 complex with a fixed number of water molecules, ten 100 ps
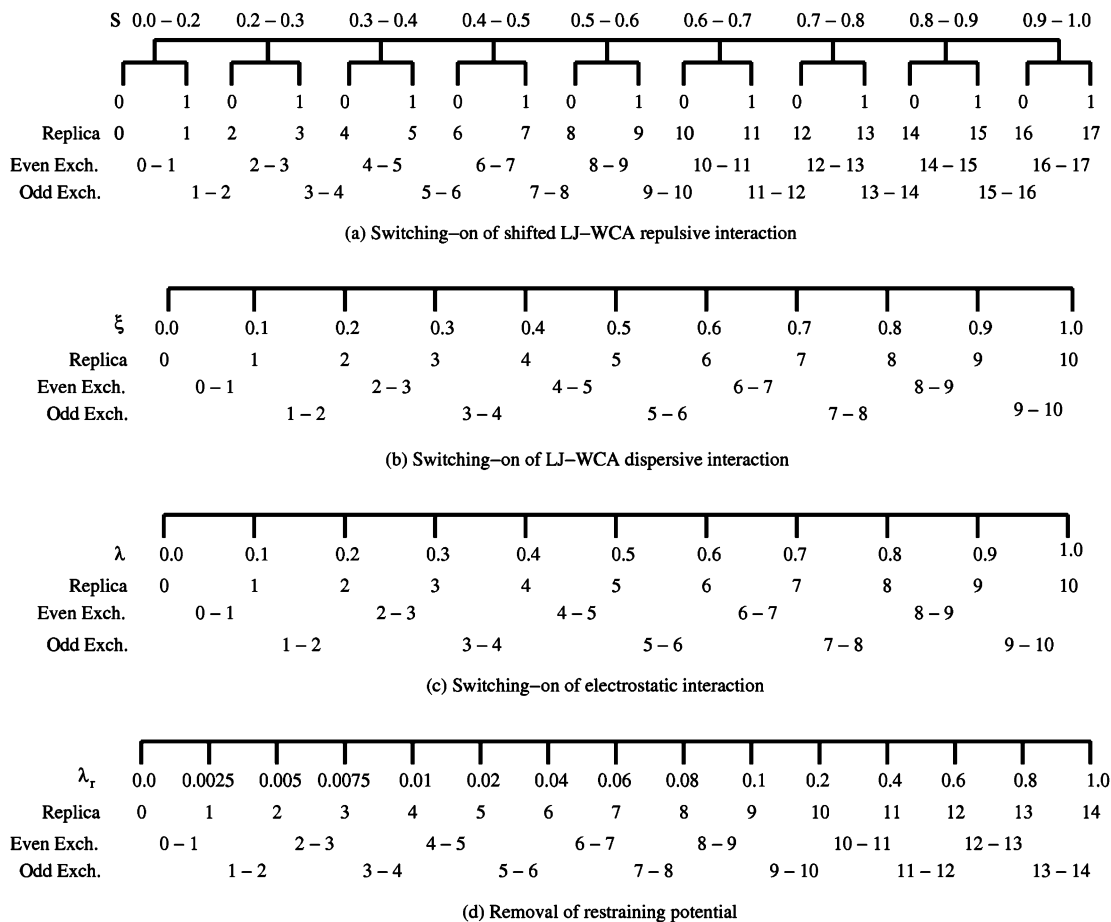
Letter

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2585**

Figure 2. REPDSTR implementation of the staging simulation protocol. Note that the LJ-WCA repulsive interaction is decomposed into 9 stages where each stage consists of two windows with 0 and 1 denoting the initial and final FEP state, respectively. All the windows of the same type were run in a single parallelized job. Each FEP window has its corresponding replica label, and each replica further occupies multiple processors, of which only the root (I/O) process (vertical solid line) is shown. Replica-exchange attempts occur between the two root processes of a selected replica pair.

Table 1. Hydration Free Energy and the Individual Components for Water

MD run	exchange frequency	ΔG_{rep}	ΔG_{disp}	ΔG_{elec}	ΔG	expt
40 ps	0	4.79 ± 0.11	-2.81 ± 0.03	-8.09 ± 0.07	-6.12 ± 0.14	-6.3
	1 /1000 steps	5.10 ± 0.16	-2.87 ± 0.01	-8.20 ± 0.12	-5.97 ± 0.23	
	1/100 steps	5.11 ± 0.15	-2.87 ± 0.02	-8.13 ± 0.08	-5.89 ± 0.18	
100 ps	0	5.12 ± 0.10	-2.88 ± 0.01	-8.20 ± 0.05	-5.95 ± 0.11	
	1/1000 steps	5.11 ± 0.06	-2.87 ± 0.01	-8.21 ± 0.07	-5.97 ± 0.12	
	1/100 steps	5.09 ± 0.07	-2.88 ± 0.01	-8.21 ± 0.06	-6.00 ± 0.12	

production runs were performed for the binding site (with GSBP) and hydration, respectively, with a replica-exchange frequency of 1/100 steps. The systems were propagated with a 2 fs time step using Langevin dynamics. The GCMC simulation algorithm, which allows the number of water molecules to fluctuate in equilibrium with an infinite bulk,[4] was employed with a replica exchange frequency of 1/100 steps. The force field parameters and initial structure of camphor-P450 complex were taken from ref 4. To estimate the statistical convergence of the calculations, 10 independent FEP/REMD simulations (10 × 100 ps) were performed consecutively for each system with a fixed number of water molecules, each starting from the configuration saved at the end of the previous run. In the FEP simulations combining GCMC and REMD, 30 consecutive calculations were performed to eliminate the influence of structural relaxation on the calculated free energy.

For all calculations, the energies were collected during the production run and postprocessed using the weighted histogram analysis method (WHAM).[19] The results for the hydration free energy of water and benzene in bulk water are given in Tables 1 and 2, respectively. The results for the binding of camphor to cytochrome P450 are summarized in Tables 3 and 4.

Results and Discussions

A. Hydration Free Energy. Tables 1 and 2 report the calculated hydration free energy for water and benzene using different replica-exchange frequencies and simulation length. For each exchange frequency, an acceptance ratio of about 30% to 40% was achieved using the λ-swap moves and the present set of window parameters. A similar acceptance ratio was obtained for the absolute binding free energy calculation. A

Table 2. Hydration Free Energy and Individual Components for Benzene

MD run	exchange frequency	ΔG_{rep}	ΔG_{disp}	ΔG_{elec}	ΔG	expt
40 ps	0	13.46 ± 0.47	−12.63 ± 0.18	−1.88 ± 0.04	−1.05 ± 0.45	−0.87
	1/1000 steps	14.41 ± 0.31	−13.07 ± 0.06	−1.89 ± 0.06	−0.55 ± 0.29	
	1/100 steps	14.45 ± 0.39	−13.01 ± 0.07	−1.85 ± 0.05	−0.41 ± 0.39	
	1/10 steps	14.67 ± 0.45	−13.07 ± 0.07	−1.90 ± 0.10	−0.30 ± 0.50	
100 ps	0	14.47 ± 0.20	−13.06 ± 0.06	−1.87 ± 0.04	−0.45 ± 0.19	
	1/1000 steps	14.50 ± 0.21	−13.06 ± 0.04	−1.86 ± 0.06	−0.42 ± 0.18	
	1/100 steps	14.49 ± 0.11	−13.03 ± 0.05	−1.86 ± 0.03	−0.41 ± 0.13	
	1/10 steps	14.49 ± 0.13	−13.03 ± 0.08	−1.86 ± 0.07	−0.41 ± 0.15	

Table 3. Binding Site Free Energies of Camphor-P450 Complex with a Fixed Number of Water Molecules at Binding Site and Hydration Free Energies of Camphor

	binding site		bulk water		
	fixed number of water (MD)	fixed number of water (REMD)	MD	REMD	expt
ΔG_{rep}	7.99 ± 0.13	8.32 ± 0.29	21.15 ± 0.21	21.06 ± 0.30	...
ΔG_{disp}	−35.50 ± 0.06	−35.97 ± 0.03	−21.15 ± 0.15	−21.12 ± 0.08	...
ΔG_{elec}	−1.52 ± 0.04	−1.25 ± 0.04	−4.80 ± 0.04	−4.82 ± 0.04	...
$\Delta\Delta G_{rstr}$	9.98 ± 0.03	10.07 ± 0.04	...	...	...
sum	−19.05 ± 0.14	−18.83 ± 0.33	−4.80 ± 0.15	−4.88 ± 0.30	−3.5
ΔG_b^0	−14.25 ± 0.21	−13.95 ± 0.55	...	...	−7.75

Table 4. Binding Free Energies of Camphor to p450

	MD run	ΔG_{rep}	ΔG_{disp}	ΔG_{elec}	ΔG_b^0	expt
GCMC-REMD	run 0−9	13.63 ± 1.23	−36.10 ± 0.10	−1.21 ± 0.02	−8.73 ± 1.43	−7.75
	run 10−19	15.61 ± 0.83	−36.18 ± 0.03	−1.24 ± 0.02	−6.86 ± 0.99	
	run 20−29	15.76 ± 0.56	−36.19 ± 0.03	−1.24 ± 0.02	−6.71 ± 0.69	
GCMC-MD	run 0−9	13.58 ± 1.19	−36.00 ± 0.03	−1.30 ± 0.01	−8.87 ± 1.31	
	run 10−19	14.32 ± 0.53	−36.05 ± 0.02	−1.29 ± 0.03	−8.16 ± 0.48	
	run 20−29	15.75 ± 0.73	−36.20 ± 0.05	−1.31 ± 0.03	−6.91 ± 0.64	

higher acceptance ratio can be obtained by increasing the number of windows but that would also increase the computational cost. The statistical uncertainties (standard deviation) for the total hydration free energy and the various components were calculated from the 10 separate production runs.

For the solute water molecule, the total hydration free energy from the 100 ps production runs converges toward a value of −6.0 kcal/mol. The three individual contributions are more informative of the performance of the FEP/REMD than the total free energy, as they display different convergences. The electrostatic and dispersive components converge, consistent with the observation that switching those relatively "soft" interactions is not associated with any sudden reorganization of the surrounding bulk. In contrast, the convergence of the repulsive component is more clearly improved by replica-exchange. The trends are clearer with the 40 ps production runs. With straight FEP/MD (no replica-exchange), the value of ΔG_{rep} is 4.97 kcal/mol (based on 10 independent runs), which systematically deviates by ∼0.3 kcal/mol from the converged value of 5.09 kcal/mol. Such deviation disappears when replica-exchange is attempted every 1000 or 100 steps. This observation is consistent with the fact that extensive sampling is necessary to sample the cavity formation associated with the repulsive solute−solvent interaction. With FEP/REMD, the ΔG_{elec} calculated from the 100 ps runs is essentially unchanged compared to the 40 ps runs, though the uncertainty is slightly smaller. The two nonpolar components converge well; the averaged ΔG_{rep} and ΔG_{disp} of the 40 ps FEP/REMD runs are almost the same as those of 100 ps runs, regardless of the relatively larger uncertainty due to shorter sampling.

Benzene provides a more interesting testing ground to illustrate the gain by FEP/REMD because of its larger size. The absolute hydration free energy of benzene calculated from 10 independent 40 ps production runs without replica-exchange (−1.05 kcal/mol) deviates systematically from the best-converged value obtained by FEP/REMD (−0.41 kcal/mol). This demonstrates the enhanced convergence from FEP/REMD. With respect to the impact of increasing the replica-exchange frequency and the sampling time, the individual components display similar trends as water shown in Table 1. In the 40 ps runs, the largest deviation again comes from ΔG_{rep} and a systematic deviation of almost −1.0 kcal/mol is avoided by replica-exchange. It is also noteworthy that the contribution from van der Waals dispersion (−12.63 kcal/mol) deviates systematically from the converged value (−13.03 kcal/mol).

The statistical uncertainty in the 40 ps simulation runs exhibits a curious "turnover" with increasing exchange frequency, i.e., the statistical uncertainty is larger when replica exchanges are attempted every 10 steps (±0.45 kcal/mol) than every 1000 steps (±0.31 kcal/mol). It is likely that the cause of this is the difficulty to sample the cavity formation associated with the harsh solute−solvent repulsive interaction. With higher exchange frequency, the trajectories jump along the λ_{rep} coupling parameter rapidly, which results in larger fluctuations in the energy samples. This situation is hidden for small solutes such as a water molecule, while the larger benzene molecule requires a more important surrounding reorganization of the solvent upon solute insertion. These observations suggest that a systematic investigation of the relation between the exchange frequency and length of MD trajectory should be the object of further
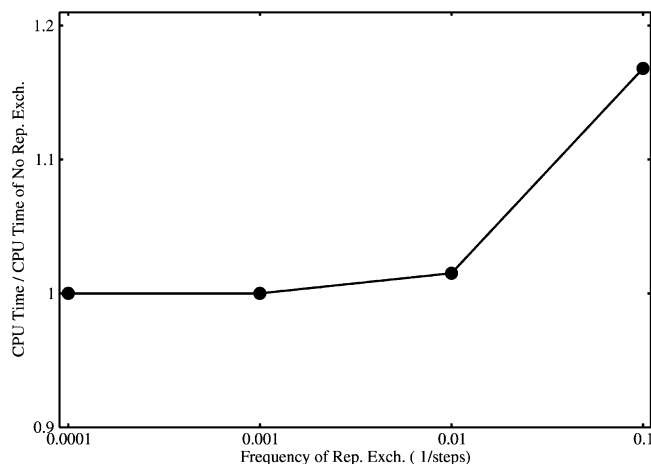
Letter

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2587**

Figure 3. Cost of replica-exchange MD simulation relative to that without replica-exchange. The CPU time of simulation without replica-exchange is denoted with exchange frequency 1/10000 steps (0.0001).

Figure 4. Binding site free energy convergence of GCMC/(RE)MD simulations. Note that GCMC/REMD shows faster convergence than GCMC/MD simulation.

analysis. Of relevance to such issues, it was recently shown that Hamiltonian replica exchange provides a direct route for minimizing the hysteresis error in hydration free energy calculations.[20]

As shown in Tables 1 and 2, the free energy with replica-exchange converges better with increased exchange frequency. To assess the additional computational cost associated with the exchanges, Figure 3 shows the CPU time of FEP/REMD relative to that of FEP simulations without replica-exchange. It can be seen that with exchange frequencies up to 1/100 steps, there is no significant increase of the CPU time, while for 1/10 steps, the CPU time increases by ~17%. Thus, an exchange frequency of ~1/100 steps seems to be an optimal choice with the current FEP/REMD implementation.

B. Binding Free Energy. The convergence of a calculation of the absolute binding free energy of camphor for p450 is challenging because the binding site is deeply buried within the protein and is not solvent accessible. For this reason, the movement of water molecules in and out of the binding pocket is extremely slowly. This is expected to complicate the convergence of FEP/MD calculations. It was shown previously that the contribution from the repulsive interaction is grossly underestimated in FEP/MD calculations initiated from the structure of the bound complex with a fixed number of water molecules in the buried site. Reasonable results are obtained if the number of water molecules is allowed to fluctuate via a GCMC algorithm.[4] It is of interest to test if replica-exchange without GCMC is able to cure this problem. To this end, the FEP/REMD calculation was first performed with a fixed number of water molecules corresponding to the X-ray structure of the bound camphor-P450 complex. In Table 3, it can be seen that REMD does not actually improve the convergence of the repulsive interaction contribution. The resulting binding free energy is too favorable by almost 5−6 kcal/mol, similar to previous results.[4] This shows that the FEP/REMD scheme, by itself, is unable to enhance the sampling of different number of water molecules in the buried binding pocket. Combining replica-exchange with GCMC and MD to perform FEP calculations appears to be necessary in this case. Table 4 gives the results for the calculations performed within a FEP/GCMC-
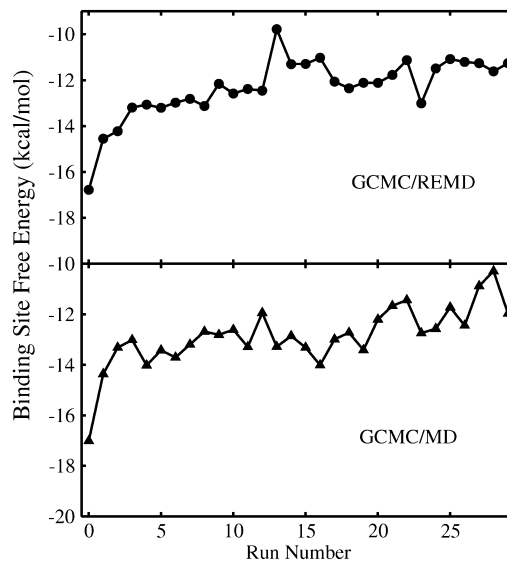
REMD scheme and compares those with the FEP/GCMC-MD. It can be seen that both methods yield a reasonable total binding free energy, indicating that the GCMC algorithm is essential to treat the buried p450 binding site. Again, the repulsive free energies estimated from FEP/GCMC-REMD converged better than that from FEP/GCMC-MD simulation. For the absolute hydration free energy of camphor in bulk water, no noticeable gain is observed with FEP/REMD compared to FEP/MD. In both approaches, the repulsive and dispersive interactions nearly cancel out. This is consistent with the previous results with water and benzene (Tables 1 and 2).

Table 4 shows that from the 10th MD run the repulsive free energies are essentially converged to 15.6 kcal/mol, whereas the corresponding results from the FEP/GCMC-MD simulation are underestimated by ~1.4 kcal/mol. Ultimately, the average repulsive free energies of both methods reach a value about 15.75 kcal/mol but only after the 20th run. Figure 4 provides a more detailed picture for the convergence of FEP/GCMC-REMD calculations. One can see that some structural relaxation dominates the early stage of the simulations with and without replica-exchange. The free energy from the FEP/GCMC-REMD simulations starts to fluctuate around a stable value of −12.0 kcal/mol from the 12th run. In contrast, the FEP/GCMC-MD calculation seems to be drifting up even after the 20th run. This shows that replica-exchange significantly accelerates the configurational sampling of the receptor−ligand−solvent complex. The calculation of the electrostatic and dispersive contributions benefits less markedly from REMD, probably because they are switched-on after the formation of a cavity and the fluctuations of the surrounding water molecules are smaller. As observed in Table 4, the dispersion free energy and charging free energy of FEP/GCMC-REMD are very close to the results from FEP/GCMC-MD. The introduction of replica-exchange does not have a large impact in the case of the geometrical restraining potential (Table 3).

Conclusion

A free energy perturbation staging protocol with replica-exchange molecular dynamics (FEP/REMD) implemented

with a distributed in a parallel/parallel mode replica strategy REPDSTR has been described and implemented for large supercomputing platforms. As a natural outcome of this implementation, replica-exchange over configuration with different thermodynamic coupling parameters was shown to enhance the sampling for the calculation of hydration free energies. The benefit is most striking in the case of the repulsive contribution to the free energy (related to the cavity formation). This would be expected to hold in general for ligand binding at the solvent-exposed site near the surface of a receptor. For the deeply buried binding site, the REMD combined with the GCMC algorithm significantly accelerates the free energy convergence. Applications to the present FEP/REMD method to produce a general protocol to different ligand binding processes are currently underway.

References

(1) Deng, Y.; Roux, B. *J. Phys. Chem.* **2004**, *108*, 16567–16576.

(2) Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

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

(4) Deng, Y.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 115103.

(5) Rick, S. W. *J. Chem. Theory Comput.* **2006**, *2*, 939–946.

(6) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84*, 775–786.

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

(8) Fajer, M.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1565–1569.

(9) Min, D.; , H.; , L.; , G.; , L.; , R.; , B.- P.; , W., Y. *J. Chem. Phys.* **2007**, *126*, 144109.

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

(11) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

(12) Mitsutake, A.; Okamoto, Y. *Phys. Rev. E* **2009**, *79*, 047701.

(13) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13711–13718.

(14) Brooks, B. R.; et al. *J. Comp. Chem.* **2009**, *30*, 1545.

(15) Woodcock, H. L., III.; Hodoscek, M.; Gilbert, A. T. B.; Gill, P. M. W.; Schaefer, H. F., III.; , R., B. B. *J. Comput. Chem.* **2007**, *28*, 1485–1502.

(16) Woodcock, H. L., III.; Hodoscek, M.; Sherwood, P.; Lee, Y.; Schaefer, H.; Brooks, B. *Theor. Chem. Acc.* **2003**, *109*, 140–148.

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

(18) Im, W.; Berneche, S.; Roux, B. *J. Chem. Phys.* **2000**, *114*, 2924–2937.

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

(20) Wyczalkowski, M. A.; Pappu, R. V. *Physical Review* **2008**, *77*, 026104.

CT900223Z

JCTC Journal of Chemical Theory and Computation

Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials

Luca Maragliano,[†] Eric Vanden-Eijnden,*,[‡] and Benoît Roux*,[¶],[†]

Department of Biochemistry and Molecular Biology, University of Chicago, 5801 South Ellis Avenue, Chicago, Illinois 60637, Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012, USA, and Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, USA

Abstract: Milestoning is a method aimed at reconstructing the statistical properties of the long-time dynamics of a system by exploiting the crossing statistics of a set of hypersurfaces, called the "milestones", placed along the reaction coordinate [Faradjian and Elber, *J. Chem. Phys.* **2004**, *120*, 10880]. Recently, Vanden-Eijnden and Venturoli [*J. Chem. Phys.* **2009**, *130*, 194101] showed that when a complete Voronoi tessellation of the configurational space is available, milestoning can be reformulated to utilize the statistics from a series of independent simulations, each confined within a single cell via strict reflections at the boundaries. As a byproduct, this "Voronoi tessellated milestoning" method also permits to compute the free energy of the tessellation. Here, the method is extended to support the usage of differentiable restraining potentials to confine the trajectories within each cell.

1. Introduction

Conformational changes in large biomolecules are complex and slow processes taking place on very long time scales, which often extend well beyond the reach of brute force molecular dynamics (MD) simulations. An emerging class of techniques attacks this problem by first trying to determine an optimal reaction pathway (or pathways) for the transition in a space of very high dimensionality, without making any a priori assumptions about the mechanism.[1−5] Once such an optimal pathway has been identified, a strategy then consists of inferring the statistical properties of the long-time dynamics by exploiting the information harvested from relatively short independent trajectories.[6−9] Of particular interest is the milestoning method introduced by Faradjian and Elber,[8] in which the dynamical properties of a reactive event are reconstructed out of a series of short trajectories between a set of hypersurfaces (the "milestones"), disposed sequentially along the reaction pathway between the reactant and the product.[10−14] Recently, Vanden-Eijnden and Venturoli extended milestoning by identifying the edges of the cells of a complete Voronoi tessellation of configurational space as the milestones.[9] It was shown that the "Voronoi tessellated milestoning" could be formulated to require only a series of independent simulations, each confined within a single Voronoi cell, leading to a simplification and an increase in robustness of the original algorithm. The confinement was realized by a simple collision rule via strict hard-wall reflections (velocity inversion) at the boundaries.[4]

Despite its formal simplicity, the hard-wall boundary condition involves modifications of the dynamical propagation algorithm at the heart of MD source codes. This makes its implementation in widely used biomolecular simulation MD packages (e.g., AMBER,[15] CHARMM,[16] GROMACS,[17] and NAMD[18]) somewhat cumbersome, and may even affect performance. With the aim of extending the range of applicability of the method, in this work, we propose and test an alternative strategy based on the introduction of continuous and differentiable restraining potentials to confine the system within a given cell. We illustrate how the formalism presented in ref 9 can still be used, with minor modifications, to compute the equilibrium probability and the rates of transitions for the original dynamics. The possibility to use potentials will facilitate the application of the Voronoi tessellated milestoning because the introduction of user-defined external forces is straightfoward. We note that the confinement with potentials can also affect performance because portions of trajectories that are transiently out of the cells are discarded in the analysis, but this effect can be minimized by proper tuning of the parameters in the potentials.

In the remaining part of this letter, we first describe the details of the restraining potentials and the key quantities to be computed for free energy and milestoning calculations, and then

* Corresponding authors. E-mail: eve2@cims.nyu.edu (E.V.-E.); roux@uchicago.edu (B.R.).
† University of Chicago.
‡ New York University.
¶ Argonne National Laboratory.

we illustrate the method by applying it to two simple examples, a numerical toy model and a solvated alanine dipeptide.

2. Theoretical Developments

2.1. Soft Wall Restraining Potentials. Let us consider a molecular system with coordinates x and potential energy $U(x)$. In most biomolecular applications, the dimensionality of the system is very high, but most of the variables can be unimportant for the description of a reactive mechanism. For this reason, it is customary to introduce a smaller number of collective variables to characterize the process. Let us indicate the set of collective variables, functions of x, by $\{z(x)\}$. Suppose we are given a set of K points in z-space, $(z_1, z_2, ..., z_K)$, that we call centroids. These support a complete and unique Voronoi tessellation of the original Cartesian space of the system, where each cell α is defined such that $\|z(x) - z_\alpha\| < \|z(x) - z_\gamma\|$ for all $\gamma \neq \alpha$; $\|\cdot\|$ indicates the norm in some metric, which shall be assumed Euclidean for the sake of simplicity.

It was shown previously,[9] via a reformulation of the milestoning algorithm,[8] that the statistical properties of the long-time dynamics of the system can be reconstructed from independent simulations, each confined inside one of the K cells. The key feature required for this result is that the confined dynamics must be equivalent to that from a long unbiased trajectory passing through the same set of cells.[4,9] More specifically, the confinement must leave unperturbed the dynamical properties of the systems when it is in the interior of the cell as well as the probability flux in and out of the cells. In ref 9, the confinement was realized with a strict reflection rule,[4] i.e., by reversing the velocity of the atoms when a trajectory attempted to cross the hyperplane between two neighboring cells α and γ. An alternative strategy, explored here, is to introduce planar half-pseudoharmonic restraining potentials to confine the system within each cell:

$$u_\alpha(x) = \sum_{\gamma=1}^{K} \frac{k_w}{2} (\hat{\mathbf{n}}_{\alpha\gamma} \cdot [z(x) - \bar{z}_{\alpha\gamma}])^2 \, H(\hat{\mathbf{n}}_{\alpha\gamma} \cdot [z(x) - \bar{z}_{\alpha\gamma}])$$

(1)

where the unit vector $\hat{\mathbf{n}}_{\alpha\gamma} = (z_\gamma - z_\alpha)/\|z_\gamma - z_\alpha\|$ is normal to the hyperplane separating the cells α and γ, $\bar{z}_{\alpha\gamma} = (z_\gamma + z_\alpha)/2$ is the midpoint between the centroids α and γ, and H is the Heaviside step function. The real parameter k_w determines the strength of the restraining potential. Hereafter, we will refer to the present approach as the *soft walls* (SW) restraints, to emphasize the distinction with the strict *hard walls* (HW) reflections used in refs 9 and 4.

An illustration of $u_\alpha(x)$ is given in Figure 1. The restraining potential $u_\alpha(x)$ is 0 in the interior of the cell α, and it acts as a penalty function toward the cell boundary when the trajectory crosses any of the hyperplanes separating α from other cells.

It can be understood from the following argument that the fundamental requirements of the confined dynamics can be met when the restraints of eq 1 are used. First, since $u_\alpha(x) = 0$ inside the cell, using eq 1 guarantees that the equilibrium distribution inside the cell is the same as that of an unbiased simulation, apart for small errors at the boundaries related to time discretization. Moreover, this implies that we also have the correct fluxes in and out of the cell (otherwise, their effect would
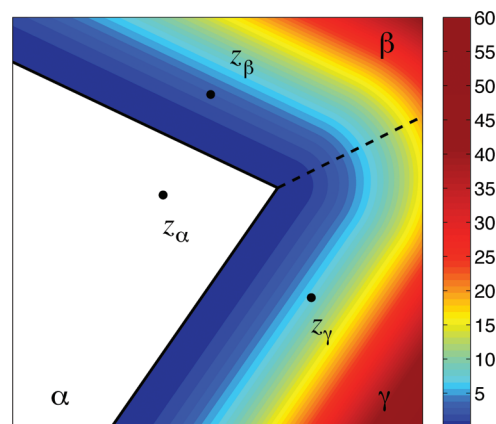

Figure 1. An illustration of the potential $u_\alpha(x)$ used to constraint a trajectory in cell α, eq 1. Portions of three cells α, β, and γ with centroids $(z_\alpha, z_\beta, z_\gamma)$ are shown. The potential is 0 in cell α (white region), and its units are arbitrary. Cell edges are represented as black lines. Note that the edge between cells β and γ (dashed black line) has no effect on the restraint, and it is shown only for the purpose of clarity.

propagate in and spoil the distribution inside), and hence, the pieces of trajectory of the restricted simulation are indistinguishable from the pieces of an unbiased trajectory passing through the same cell.

It is possible to compute the equilibrium probability and the rates of transitions for the original dynamics from the simulations confined within the Voronoi cells by the HW reflections.[9] In the following sections, we show that, with minor modifications, the formalism can still be used with the SW restraints.

2.2. Free Energy of the Tessellation. The conservation of probability flux through the boundaries of the cell gives a way to compute, π_α, the equilibrium probability of the system to be in cell α, and the associated free energy, $G_\alpha = -\beta^{-1}\log \pi_\alpha$ (where $\beta = 1/k_B T$, with k_B the Boltzmann constant and T the temperature), as summarized hereafter.[4,9] The rate of escape from cell α to cell γ, conditional on the system being in the cell α, is defined as $\nu_{\alpha\gamma} = N_{\alpha\gamma}/T_\alpha$, where $N_{\alpha\gamma}$ is the number of collisions with the boundary separating the cells α and γ, and T_α is the total simulation time spent inside cell α. The normalized equilibrium probability π_α solves the following equations involving the rates of escape $\nu_{\alpha\gamma}$:

$$\sum_{\substack{\gamma=1 \\ \gamma\neq\alpha}}^{K} \pi_\gamma \nu_{\gamma\alpha} = \sum_{\substack{\gamma=1 \\ \gamma\neq\alpha}}^{K} \pi_\alpha \nu_{\alpha\gamma}, \qquad \sum_{\alpha=1}^{K} \pi_\alpha = 1 \qquad (2)$$

This equation expresses that, at statistical steady state, the total probability flux in cell α must be equal to the flux out of α by conservation of the total probability. In ref 9, T_α was simply the total simulation time with the system confined in cell α by means of HW. With SW, eq 2 is still valid, as long as one counts only the portion of trajectory spent inside the cell α as the time T_α.

2.3. Milestoning. Identifying the edges of the Voronoi tessellation as milestones, the dynamics of the system is reduced to that of a discrete state continuous-time Markov chain in the state space of the milestones indices.[9] By indexing the milestones as i and j, this amounts to defining a rate matrix q_{ij}, whose elements are given by

$$q_{ij} = N_{ij}/R_i \qquad (3)$$

The factors N_{ij} and R_i in this equation can be expressed[9] in terms of the average properties extracted from simulations confined to the cell α, weighted by the equilibrium probability π_α of finding the system in cell α:

$$N_{ij} = \sum_{\alpha=1}^{N} \pi_\alpha (N_{ij}^\alpha/T_\alpha), \qquad R_i = \sum_{\alpha=1}^{N} \pi_\alpha (R_i^\alpha/T_\alpha) \quad (4)$$

Here T_α is the duration of the simulation confined in cell α, N_{ij}^α is the number of transitions from edges i to j observed during this simulation, and R_i^α is the total time that edge i was the last edge to be hit during this simulation (i.e., $\sum_i R_i^\alpha = T_\alpha$). The only requirement to use eqs 4 with SW is to prune the trajectory to its portion that is strictly inside the cell α before computing N_{ij}^α, R_i^α, and T_α (i.e., the parts spent outside the cell in the SW region must be discarded).

The rate matrix q_{ij} specifies completely the dynamics of the Markov process, and hence, it can be used to compute many important quantities as, for example, the mean first passage times (MFPTs) from any milestone to any other.[9,14] For instance, if τ_i^N with $i = 1, ..., N-1$ denote the MFPTs from milestone i to milestone N ($\tau_N^N = 0$ by definition), then these MFPTs can be computed by solving the linear system of equations:

$$\sum_{j=1}^{N-1} q_{ij}\tau_j^N = -1, \qquad i = 1, ..., N-1 \qquad (5)$$

3. Results and Discussion

In the following, we illustrate the implementation of milestoning with SW restraints on two simple systems, the Mueller potential[19] and the solvated alanine dipeptide. Prior to this, an important point concerning the position of the centroids supporting the Voronoi tessellation deserves special attention. While eqs 3−5 can be applied to any set of hypersurfaces, it has been shown that the formalism gives exact MFPTs, if the hypersurfaces used as milestones are chosen as the isocommittor surfaces of the reaction.[14] From a practical viewpoint, the isocommittor surfaces can be calculated approximately by the string method and its variants.[3−5,20−22] In the string method, the transition path is represented as by an ordered sequence of K discrete "images", $\{z^{(1)}, z^{(2)}, ..., z^{(K)}\}$, in the space of collective variables $z(x)$. This suggests[9] to first use the string method to determine the transition path by optimizing the position of the K images and then use these images as centroids to support the Voronoi tessellation. Naturally, the edges between the cells are hyperplanes perpendicular to the optimized path and, hence, approximations of the isocommittor surfaces.

3.1. The Mueller Potential. We first consider a simple 2-dimensional system evolving on the Mueller potential energy surface according to Langevin dynamics (i.e., we take $\{z(x,y)\} \equiv (x,y)$, the coordinates of the system). The same model was studied in ref , so for comparison, we took the same centroids as generators of the Voronoi tessellation. These are 18 equidistant images (the red circles in Figure 2), computed with the finite-temperature string method.[4] All simulations were performed by integrating Langevin equations of motion with the second-order algorithm of ref 23 by taking $\beta^{-1} = 20$, friction
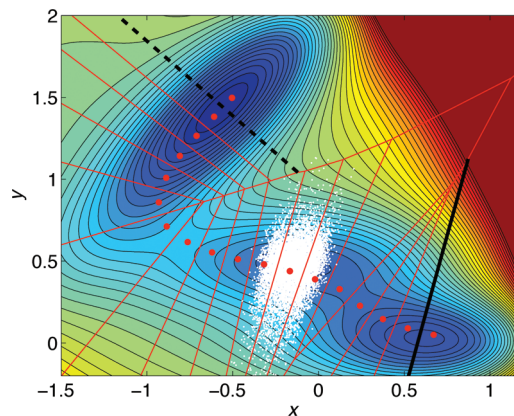

Figure 2. Mueller potential with a set of milestones, corresponding to the edges of Voronoi cells, generated by images along the converged string from the finite-temperature string method. The successive positions of the system at every time step along a piece of simulation with SW are also shown as white dots. MFPTs, discussed in the text, are calculated from the dashed to the continuous black line.
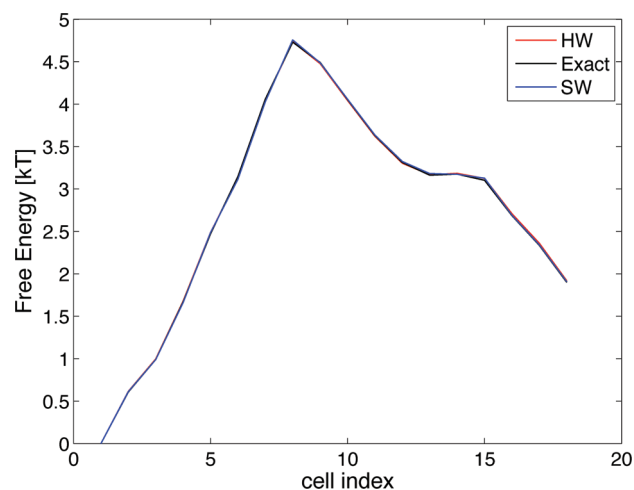

Figure 3. Free energy $G_\alpha = -\beta^{-1}\log \pi_\alpha$ of the Voronoi tessellation from HW and SW simulations on the Mueller potential compared with the exact one obtained by numerical integration. Numbering of the cells goes from top-left to bottom-right in Figure 2.

400, and a time step $\Delta t = 10^{-4}$. Hard and soft wall trajectories of 10^8 steps were generated in each cell.

Figure 2 shows the Mueller potential together with the Voronoi tessellation associated to the images from the string. The edges of the cells were taken as milestones. The successive positions of the system at every time step along a piece of simulation with SW are also shown as white dots. Note that, as already pointed out before, the portion of trajectory inside the cell samples the equilibrium distribution associated to the underlying potential in the cell.

Figure 3 shows the free energy $G_\alpha = -\beta^{-1}\log \pi_\alpha$, where π_α is the solution of eq 2, corresponding to the tessellation shown in Figure 2. Results are presented for SW and HW simulations and numerical integration (Exact). The numbering of the cells goes from 1 to 18 from top-left to bottom-right in Figure 2. Table 1 summarizes results for MFPTs calculations considering the transition from the black dashed line to the black continuous line in Figure 2. The table also shows the mean and the extremal

Table 1. Kinetics for the Mueller Potential for the Transition from the Black Dashed Milestone to the Black Continuous Milestone in Figure 2[a]

	MFPT	$\langle R_i \rangle$	$R_i$max	$R_i$min	$\langle N_{ij} \rangle$	N_{ij}max	N_{ij}min
soft walls	547 ± 24	0.0065	0.4644	1.55×10^{-8}	1.35×10^{-6}	3.32×10^{-5}	6.05×10^{-11}
hard walls	553 ± 14	0.0065	0.4645	9.99×10^{-9}	1.37×10^{-6}	3.27×10^{-5}	4.10×10^{-11}
free simulation	562 ± 16	0.0065	0.4645	1.89×10^{-8}	1.37×10^{-6}	3.19×10^{-5}	4.02×10^{-11}

[a] The agreement of all quantities shows that the dynamical properties of the trajectory inside the cells are left unperturbed by the two different confinement methods, i.e., it is as if the confinement was absent. Errors on MFPTs were obtained from five different runs.
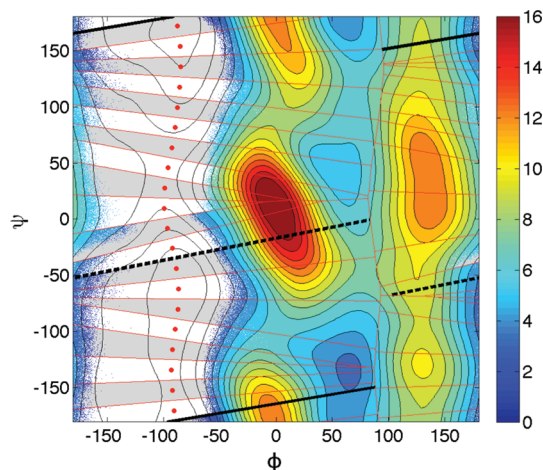

Figure 4. Backbone dihedral free energy surface of solvated alanine dipeptide (units are kcal/mol) with a set of milestones (red lines) corresponding to the edges of Voronoi cells generated by images along the converged minimum free energy path (red circles) and the helical to extended transition. The successive values of dihedral angles at every time step along pieces of simulations with HW are shown as white and gray dots. MFPTs, discussed in the text, are calculated from the dashed to the continuous black line.

Figure 5. Free energy of the Voronoi tessellation from HW and SW sampling for the solvated alanine dipeptide. Numbering of the cells goes from 1 to 20, counting from top to bottom in Figure 4.

Figure 6. Rates of escape from cell #2 (i.e., associated to the second string image counting from top to bottom in Figure 4), ν_{21}, top line, and ν_{23}, bottom line, as a function of the trajectory length for the alanine dipeptide SW simulation with $k_w = 200$ kcal/mol/rads^2.

values of the elements needed for the calculation of the q_{ij} matrix defined in eq 3, computed from SW and HW simulations and a long free trajectory (2×10^9 steps). The excellent agreement shows that the confinements do not introduce biases in the calculations of q_{ij}.

3.2. Solvated Alanine Dipeptide. In order to test the implementation of SW and HW in a realistic biomolecular context, we examine the transition from helical to extended conformation of the alanine dipeptide solvated in explicit water. We describe the transition here by using two collective variables, the ϕ and ψ backbone dihedral angles, neglecting the role of the solvent degrees of freedom.[24] We used the CHARMM[16] code for all calculations with CHARMM22 all-atom force field[25] and the TIP3P model[26] for water molecules.

The free energy landscape in ϕ and ψ variables was computed with the single-sweep method,[27] and it is shown for reference in Figure 4 (energies are in kcal/mol). A minimum free energy path in (ϕ,ψ) space was computed using the string method.[3] 20 images along this path are represented as circles in Figure 4, and they were used as centroids for the Voronoi tessellation. Note that in our calculation, we focus on the reaction channel with $\phi < 0$, where the main metastable states are. In principle, we could also investigate the kinetics of transitions between the states with $\phi > 0$, but these are less often visited, and hence, their existence does not affect much the rates we compute. The edges of the Voronoi cells are represented as red lines in Figure
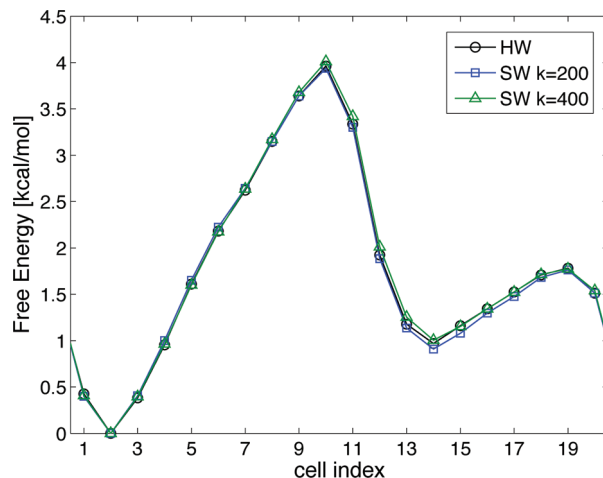
4 (taking into account periodicity). The dashed and continuous black lines are the start and the ending milestones for MFPTs calculation. The HW confinement condition was implemented into the Nosé–Hoover (NH) dynamics subroutine of CHARMM.[28] The successive values of dihedral angles at every time step along pieces of simulations with ahrd walls are shown as white and gray dots. NH dynamics was also used for SW simulations. In this case, the forces on atoms coming from the potential (1) were implemented and added to the standard CHARMM forces.

Letter

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2593**

Table 2. Kinetics for the Solvated Alanine Dipeptide for the Transition from the Black Dashed Milestone to the Black Continuous Milestone in Figure 4

	MFPT (ps)	$\langle R_i \rangle$	$R_i$max	$R_i$min	$\langle N_{ij} \rangle$	N_{ij}max	N_{ij}min
soft walls	28.33	0.0053	0.3242	6.65×10^{-7}	9.24×10^{-5}	6.87×10^{-4}	2.99×10^{-9}
hard walls	28.13	0.0053	0.3306	5.00×10^{-7}	8.86×10^{-5}	7.02×10^{-4}	1.97×10^{-8}

Figure 5 shows the free energy corresponding to the tessellation shown in Figure 4. Results are for HW and SW simulations (with two different values of the penalty constant k_w). In all cases, the trajectories in each cell were 1 ns long. Numbering of the cells goes from 1 to 20 and top to bottom in Figure 4. In order to assess convergence of the calculation, we monitored, during the simulations, the values of the escape rates from the cells, $\nu_{\alpha\gamma}$. Figure 6 shows the rates from cell two versus the length of the entire confined trajectory for SW simulations with $k_w = 200$ kcal/mol/rads^2. Escape rates from other cells converge on the same time scale.

MFPTs were computed considering the transition from the black dashed to the black continuous line in Figure 4. Results are summarized in Table 2, together with the mean and the extremal values of the elements needed for the calculation of the q_{ij} matrix defined in eq 3. The excellent agreement shows again that the confinements do not introduce biases in the calculations of q_{ij}.

The MFPT for the transition from dashed to continuous milestone in Figure 4 estimated from milestoning with HW and SW simulations is about 28 ps. For comparison, the MFPT between the same two milestones was computed by direct counting of consecutive hittings, along a free, unbiased simulation of 80 ns, obtaining 29.97 ps. These values agree with results from other CHARMM calculations of solvated alanine dipeptide in a similar simulation setup, where the MFPT for the same transition was estimated to be of about 30 ps.[29,30]

4. Concluding Remarks

Milestoning with Voronoi tessellation[9] is a method to reconstruct the dynamical properties of complex reactive systems by matching together informations obtained from multiple trajectories, each confined in a different cell of a Voronoi tessellation of configurational space. In this letter, we have illustrated how the formalism can still be applied, with minor modifications, when the confinement in the cells is realized via continuous and differentiable potentials. With respect to the original formulation, where the confinement is realized with strict reflections (velocity inversion at the boundary), the possibility to use potentials will facilitate the application of the method because user-defined external forces are easy to introduce in most MD codes and do not require modification of the dynamical propagators. However, the confinement with potentials can also affect performance because portions of trajectories that are transiently out of the cells are discarded in the analysis. Note also that we did not optimize for efficiency here, which would require adjusting the penalty constant k_w in eq 1.

As a final remark, let us point out that the method proposed in this letter, like the one in ref 9, can be trivially distributed on multiple processing nodes, since it is based on independent simulations with no required communication among them.

References

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

(2) Jonsson, H.; Mills, G.; Jacobsen, K. W. *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific: Singapore, 1998; p 385.

(3) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106.

(4) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194103.

(5) Pan, A. C.; Sezer, D.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3432.

(6) Pan, A. C.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 064107.

(7) van Erp, T. S.; Bolhuis, P. G. *J. Comput. Phys.* **2005**, *205*, 157.

(8) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.

(9) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194101.

(10) Shalloway, D.; Faradjian, A. K. *J. Chem. Phys.* **2006**, *124*, 054112.

(11) West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104.

(12) Elber, R. *Biophys. J.* **2007**, *92*, L85.

(13) Kuczera, K.; Jas, G. S.; Elber, R. *J. Phys. Chem. A* **2009**, *113*, 7461.

(14) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.

(15) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.

(16) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545.

(17) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701.

(18) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781.

(19) Mueller, K. *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 1.

(20) Vanden-Eijnden, E. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*, Springer-Verlag, Heidelberg, Volume *1*, 2007.

(21) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Chem. Phys.* **2007**, *126*, 164103.

(22) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Phys. Chem. B* **2005**, *109*, 6688.

2594 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Letter

(23) Vanden-Eijnden, E.; Ciccotti, G. *Chem. Phys. Lett.* **2006**, *429*, 310.

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

(25) Mackerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586.

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

(27) Maragliano, L.; Vanden-Eijnden, E. *J. Chem. Phys.* **2008**, *128*, 184110.

(28) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.

(29) Strodel, B.; Wales, D. J. *Chem. Phys. Lett.* **2008**, *466*, 105.

(30) Zhang, Y.; Pastor, R. W. *Mol. Simul.* **1994**, *13*, 25.

CT900279Z

JCTC Journal of Chemical Theory and Computation

Automated Event Detection and Activity Monitoring in Long Molecular Dynamics Simulations

Willy Wriggers, Kate A. Stafford, Yibing Shan, Stefano Piana, Paul Maragakis,
Kresten Lindorff-Larsen, Patrick J. Miller, Justin Gullingsrud, Charles A. Rendleman,
Michael P. Eastwood, Ron O. Dror, and David E. Shaw*

D. E. Shaw Research, New York, New York 10036

Received May 7, 2009

Abstract: Events of scientific interest in molecular dynamics (MD) simulations, including conformational changes, folding transitions, and translocations of ligands and reaction products, often correspond to high-level structural rearrangements that alter contacts between molecules or among different parts of a molecule. Due to advances in computer architecture and software, MD trajectories representing such structure-changing events have become easier to generate, but the length of these trajectories poses a challenge to scientific interpretation and analysis. In this paper, we present automated methods for the detection of potentially important structure-changing events in long MD trajectories. In contrast with traditional tools for the analysis of such trajectories, our methods provide a detailed report of broken and formed contacts that aids in the identification of specific time-dependent side-chain interactions. Our approach employs a coarse-grained representation of amino acid side chains, a contact metric based on higher order generalizations of Delaunay tetrahedralization, techniques for detecting significant shifts in the resulting contact time series, and a new kernel-based measure of contact alteration activity. The analysis methods we describe are incorporated in a newly developed package, called *TimeScapes*, which is freely available and compatible with trajectories generated by a variety of popular MD programs. Tests based on actual microsecond time scale simulations demonstrate that the package can be used to efficiently detect and characterize important conformational changes in realistic protein systems.

1. Introduction

As progress in computer technology has extended the reach of molecular dynamics (MD) simulations[1,2] from picoseconds to nanoseconds and microseconds, complex and functionally important biomolecular motions, such as protein folding and ligand binding, have become more accessible, but the resulting data sets have become increasingly large and unwieldy. Routine MD simulations currently generate trajectories consisting of thousands or millions of frames, rendering both visual inspection and data analysis difficult and time-consuming. We expect that, over time, the analysis

of these trajectories will require increasing automation, with human intervention limited to selected events of scientific interest.

We are particularly interested in the detection of significant secondary or tertiary structure rearrangements of proteins, as these motions are often of functional importance. Examples of such large-scale motions include allosteric conformational transitions and folding processes, which give rise to substantial alterations in the interactions between amino acid residues. To shed light on such phenomena, the work described in this paper focuses largely on the automated recognition of significant amino acid contact changes in MD trajectories and on measurement of the *activity*, the total number of such changes per unit time.

Our approach makes use of a particular type of "coarse-grained" model to reduce the level of detail in the spatial

---

* Corresponding author phone: (212) 478-0260; fax: (212) 845-1286; e-mail: David.Shaw@DEShawResearch.com. David E. Shaw is also with the Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032.

representations of long MD trajectories. In particular, we employ a coarse-grained model based on side chains, which offers certain advantages over models based on α-carbon atoms in the context of the present application. Three-dimensional protein structures are described using a distance matrix representation[3,4] that records all pairwise distances between the coarse-grained side chains. In contrast with traditional methods based on the use of global root mean square deviation (rmsd) measurements, the use of distance matrices does not require the translational and rotational alignment of protein structures and facilitates the identification of local structural differences.[5] Our approach decomposes structural changes into a set of key side-chain motions, providing greater sensitivity to a wide range of significant conformational changes than is typically obtained from traditional rmsd-based metrics.

We introduce two alternative approaches to identifying time-dependent contact graphs from distance matrices: a method based on distance cutoffs, which proves useful for detecting local contact formation and breaking activities, and an approach based on Delaunay tetrahedralization, which is better suited to the detection of global folding activities. A recrossing filter is used to eliminate transiently appearing or disappearing edges in the contact graph that are likely to represent random fluctuations and not biologically significant conformational changes.

In the remainder of this paper, we describe the essential elements of our approach, using four microsecond-scale simulations for illustrative purposes. For each trajectory frame, we construct a graph representing all contacts between amino acid side chains, computed using a spatially coarse-grained representation. We track changes in this graph over time, employing a median filter and a recrossing filter for the counting of discrete events that are reflected in the time-dependent contact graph. Finally, we use a kernel measure to derive activity levels from the event data. Although in this work we only examine protein trajectories, it should be relatively straightforward to generalize our approach to, for example, nucleic acids or carbohydrates.

2. Methods

2.1. Molecular Dynamics Simulations. We applied our algorithms to four all-atom MD trajectories, each approximately 1 μs in length. (More detailed system parameters are given in the Supporting Information.) Using the traditional metric of α-carbon rms deviation from the known atomic structure, Figure 1 shows distinct dynamic behavior among the chosen trajectories, which we found particularly useful for method validation. Trajectory 1 (blue) results from a 0.52 μs simulation of Src kinase. In this "generic" trajectory, the system experiences a series of conformational changes, forcing it to increasingly higher rms deviations of up to 4 Å. The "stationary" trajectory 2 (red) corresponds to a stable 1.0 μs simulation of the fast-folding triple mutant K65(NLE), N68H, K70(NLE) of chicken villin subdomain HP-35,[6] where the system remains close to the initial conformation (rms deviation ~1 Å) over the entire length of the simulation. The "diffusive" trajectory 3 (black) shows
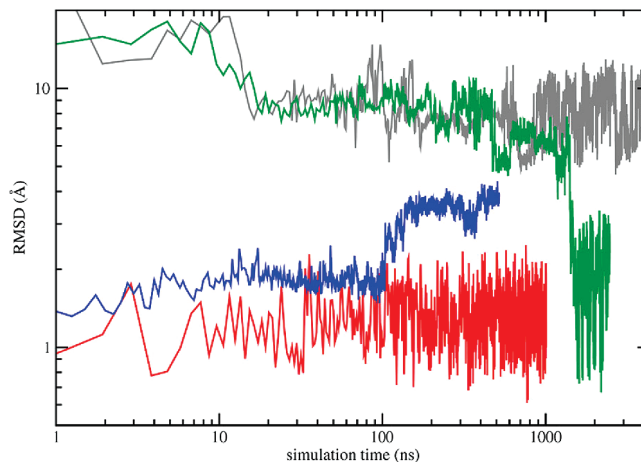

Figure 1. α-Carbon rms deviation from the native conformation as a function of simulation time: trajectory 1 (blue, Src kinase); trajectory 2 (red, villin near-native); trajectory 3 (gray, villin unfolded); trajectory 4 (green, villin folding). The double logarithmic plot bridges between the various time and spatial scales explored by the four trajectories.

the opposite behavior. Starting from an extended (unfolded) villin chain, the system remains far from the native structure during the full 4.3 μs simulation time, visiting various unfolded conformations. Finally, trajectory 4 (green) corresponds to a 2.5 μs "folding" simulation of villin. Together, the four trajectories in Figure 1 cover several scenarios that are commonly encountered in MD simulations of folded and unfolded proteins.

2.2. Coarse-Graining of Side-Chain Contacts. Noncovalent interactions between side chains, such as hydrogen bonds or salt bridges, play a critical role in protein dynamics. Singh and Thornton have shown that each of the 400 possible amino acid side chain pairings exhibits a pronounced peak in its separation histogram at a distance of 5–8 Å.[7] Following this finding, we identify a *representative atom* in each side chain for an efficient calculation of such contact separation distances. (Some side chains have more than one functional group, but our current Delaunay tetrahedralization approach relies on the choice of one representative atom per side chain.) For most residues, we define the second heavy (nonhydrogen) atom counted from the end of the chain as the representative atom. This rule takes into account the fact that in branched residues (e.g., Gln, Asp, or Arg) the end of the chain may be ambiguous, whereas the second heavy atom is straightforward to define in 14 amino acids. Of the remaining six, three aromatic residues (His, Phe, and Tyr) form special cases due to the presence of an aromatic ring; here we pick the atom at the base of the ring (closest to the main chain) as representative. The rare Trp is represented by the epsilon-2 carbon at the center of the double ring. Finally, the achiral Gly and cyclic Pro do not have extended side chains. We represent them by the α- and γ-carbons, respectively, to account for all amino acids.

The idea of reducing the level of detail is not unique to our work, and a number of similar concepts have already been described.[8–13] One possibility is to consider the hydrogen bonding network[14] as a coarse representation of relevant contacts. We have decided against using hydrogen

Event Detection in MD Simulations

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2597**

bonds because they tend to be very transient in MD simulations[15] and provide too much detail; it is often sufficient to know which amino acids are interacting. Another possibility is to select the centroids of side chains[16] or the α-carbon atoms instead of the representative atoms introduced above. Due to the widely variable sizes of side chains, however, the centroids or α-carbons are imprecise markers for interactions with neighboring residues. Alternatively, one could consider the five to seven spatial contact patterns discovered by Singh and Thornton for each of the pairings of amino acids in their *Atlas of Side-Chain Interactions*,[7] but the enumeration of such patterns for every amino acid candidate pair in every trajectory frame would be much more expensive than our simple distance metric.

Given a coarse representation of the structure, an important step in our analysis is to estimate the time-dependent contact pattern (or graph) that captures interactions between representative side-chain atoms. Any such graph approximates the actual atomic interactions, so in practice we can expect some inaccuracies in the assignment of contacts. We introduce two possible approaches for identifying contact graphs with this model: the distance cutoff and the so-called generalized masked Delaunay (GMD) tetrahedralization. Each of these graph-based concepts has its unique advantages for event detection and activity monitoring. While the distance cutoff approach is more selective with respect to local proximity relationships, which is useful for tasks such as distinguishing between the formation and breaking of contacts, the GMD approach accounts for global geometric changes and offers a way to monitor the overall structural variability. For the assignment of the contact graph, it is useful to consider advantages and limitations of these concepts in more detail. Figure 2 provides a schematic overview of proximity measures in two dimensions (the generalization to three dimensions is straightforward). The initial side-chain model is depicted in Figure 2A.

2.3. Distance Cutoff. The cutoff-based metric is the most basic proximity criterion. Contacts are based on the Euclidean distance between representative atoms, and atoms closer than a given cutoff are considered in contact. Parts B and C of Figure 2 illustrate the difficulties associated with identifying contacts by a cutoff distance. If the cutoff is too short (Figure 2B), some valid contacts may be missed, producing false negatives. If the cutoff is too long (Figure 2C), too many undesired contacts are included in the graph, leading to false positives. Such redundant graph edges are typically inconsistent with the actual nearest-neighbor interactions of side chains. In practice, a compromise between these two extreme cases must be found by adjusting the distance cutoff.

The acceptable tolerance for false positives or negatives depends on the application. For example, in α-carbon-based elastic network models, which exhibit a level of detail similar to our side-chain model, the tolerance for false positives is high. Hence, long cutoff distances of 10−15 Å are typically applied in elastic networks, about twice the separation of adjacent α-carbons.[17] Ideally, however, we select in our coarse model only those contacts that correspond to atomic contacts between side chains, requiring us to use a shorter cutoff and leading to a risk of false negatives (Figure 2B) in
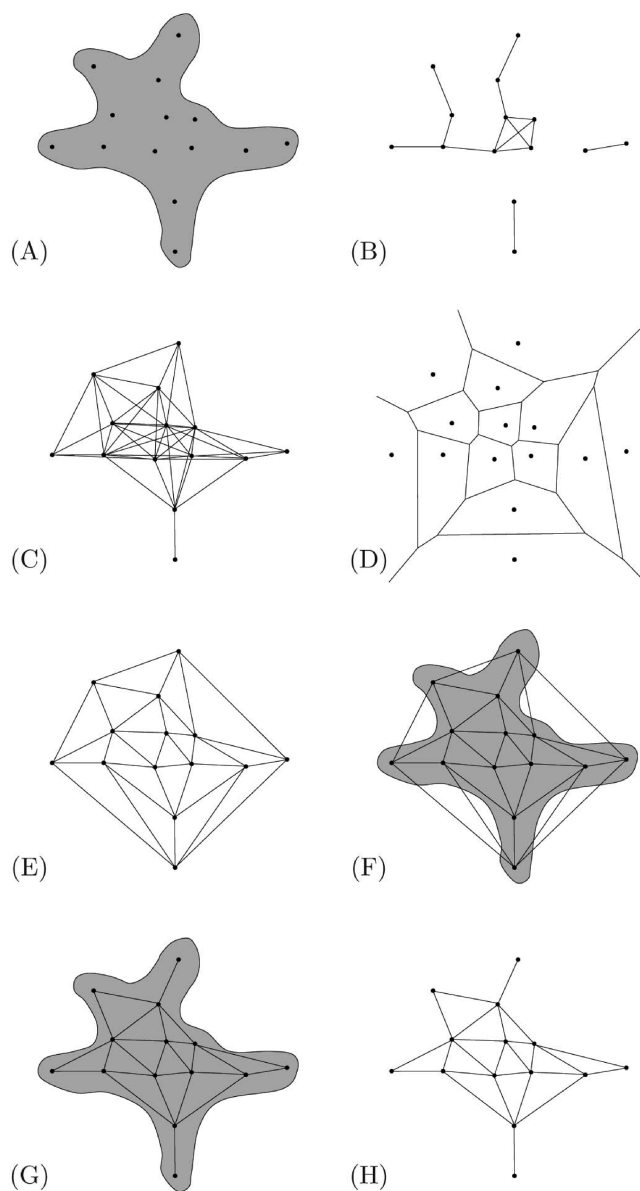
Figure 2. Idealized depiction of computational geometry concepts: (A) Coarse model (black, representative side-chain atoms) superimposed over the "protein" (gray); (B) contacts selected by distance cutoff (too short); (C) contacts selected by distance cutoff (too long); (D) Voronoi cells; (E) Delaunay triangulation; (F) Delaunay triangulation superimposed over protein; (G) masked Delaunay triangulation graph superimposed over protein; (H) masked Delaunay triangulation graph.

the resulting contact graph. The distance cutoff criterion also assumes that the side chains are densely packed and that the packing density remains invariant, which is true only for tightly folded proteins.

2.4. Generalized Masked Delaunay Tetrahedralization. The Voronoi diagram (Figure 2D) and the related Delaunay triangulation (Figure 2E) are well-known proximity measures that automatically adapt to the packing density and do not require cutoff parametrization. Voronoi cells correspond to a nearest-neighbor tessellation of the embedding space:[18] each Voronoi cell contains one representative atom (representing a single side chain) and the region of space that is closer to that representative atom than to any other. A

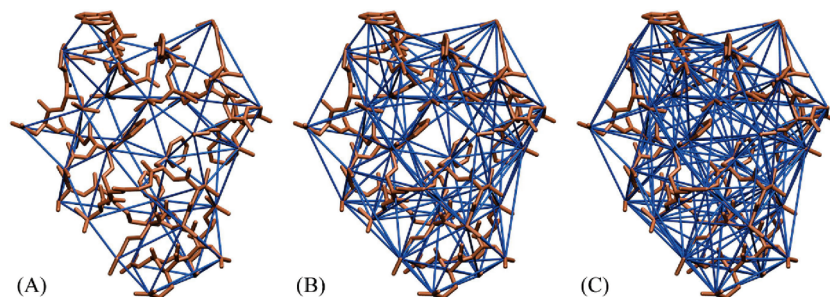**Figure 3.** General masked Delaunay (GMD) tetrahedralization of side-chain contacts (blue) in villin (brown, PDB entry 2F4K; see text): (A) Order 2 contacts; (B) order 3 contacts; (C) order 4 contacts. Molecular graphics were created with VMD.[27]

Delaunay graph is the "dual graph" of the Voronoi graph for the same set of representative atoms; one obtains the Delaunay graph by connecting representative atoms whose Voronoi cells share a face or edge. It is straightforward to generalize the first-order Voronoi cells in Figure 2D to higher order; a second-order cell, for instance, corresponds to the regions of space closest to a particular pair of representative atoms. In general, kth-order cells correspond to regions in space that are closest to a particular k-tuple.[19] Such higher order cells might be very small in size (for a depiction see Figure 2 in ref 19).

The Delaunay graph (Figure 2E) appears to be well-suited for our identification of adjacent contacts among representative atoms in three dimensions, but only a subgraph of the Delaunay graph is embedded in the protein structure (schematically shown in Figure 2F). We thus use the so-called "masked Delaunay" tetrahedralization introduced by Martinetz[20] to represent the protein shape more accurately (Figure 2G,H). The Martinetz masking algorithm takes advantage of a theorem (theorem 3 in ref 21) stating that the existence of a second-order Voronoi cell between two representative atoms is equivalent to the presence of a Delaunay edge between them (Figure 2E). An edge-defining second-order Voronoi cell is identified when it contains at least one point of a discretely sampled masking manifold. In our application, the proposed mask is the protein structure and the required discrete sampling is provided naturally by the protein atoms. Figure 3A illustrates the three-dimensional masked Delaunay tetrahedralization for villin.

The original masked Delaunay approach identifies a second-order graph (Figure 3A), connecting pairs (1-simplices) of adjacent representative atoms. We generalize the masked Delaunay approach to higher order, connecting triangles (Figure 3B), tetrahedra (Figure 3C), or, in general, $(k-1)$-simplices, where k is the order of the generalized masked Delaunay graph. This higher order generalization is motivated by the need for a discrete metric for the separation of arbitrary pairs of representative atoms in the GMD context; we use as a metric the minimum order k of the GMD graph for which the pair forms an edge. This discrete k-metric enables us to establish a recrossing filter for accurate detection of contact transitions (further discussed below). The recrossing filter aims to suppress any time-dependent spurious variations in the graph and will also suppress the effect of sampling granularity, i.e., the spacing of generic atoms in the system that might lead to missing GMD edges. To our knowledge, the GMD graph is a new concept, but the

related Voronoi cells have already been generalized to higher order, as described above.

Following Martinetz's original definition of the masked Delaunay graph,[20] and sampling the protein mask by the full atom representation, we arrive at a compact formulation of the order-k GMD as applied to biomolecular systems:

(i) Begin with the empty graph G, atom positions $\vec{v}_i \in R^3$ ($i = 1, 2, ..., N$), and representative side-chain atom positions $\vec{v}_j \in R^3$ ($j = 1, 2, ..., M$).

(ii) For each atom position $\vec{v}_i$, identify a set of k indices $S_i = \{j_1, j_2, ..., j_k\}$ and its complement S_i^C, $S_i \cup S_i^C = \{1, 2, ..., M\}$ with

$$|\vec{v}_i - \vec{v}_{j_1}| < |\vec{v}_i - \vec{v}_{j_2}| < ... < |\vec{v}_i - \vec{v}_{j_k}| < |\vec{v}_i - \vec{v}_j| \quad (j \in S_i^C)$$

(iii) Add the $(k-1)$-simplex with vertices $(\vec{w}_{j_1}, \vec{w}_{j_2}, ..., \vec{w}_{j_k})$ to G; continue with (ii) until all atoms have been explored.

For a general order k, rule (ii) implies that an edge in the GMD corresponds to a nonempty kth-order Voronoi cell, where in our case the nonempty property refers to the sampling by at least one atom in the system. The rule requires only a partial sorting of the $\vec{w}_j$, which can be efficiently implemented with complexity $O(NkM)$ per trajectory frame. The proposed GMD algorithm is efficient since it does not require an expensive geometric construction of Voronoi polyhedra or Delaunay tetrahedra.

The effect of the GMD order k on the pair distance distribution of the representative side-chain atom model is demonstrated in Figure 4. The tail of the distribution arising from the second-order GMD (a subgraph of the traditional Delaunay tetrahedralization) reaches to distances as high as 10 Å. Figure 4 shows that a 10 Å cutoff would be too permissive and would include many higher order (i.e., redundant) contacts. As a trade-off between false positives and false negatives in cutoff-based graphs, we thus recommend cutoff values of ∼7 Å, which would include the peak of the second-order GMD and only a small number of third-order GMD contacts.

2.5. Suppressing High-Frequency Motion. MD time series exhibit a considerable amount of fluctuation on short time scales, introducing noise in the conformational analysis. This noise complicates the reliable identification of significant "level shifts" in the distribution of representative atom pair distances that are relevant over longer time scales (the term "level shifts" is used in time series analysis for low-frequency changes of a nonstationary signal[22]). Such shifts are important both for the cutoff and GMD graphs since they affect
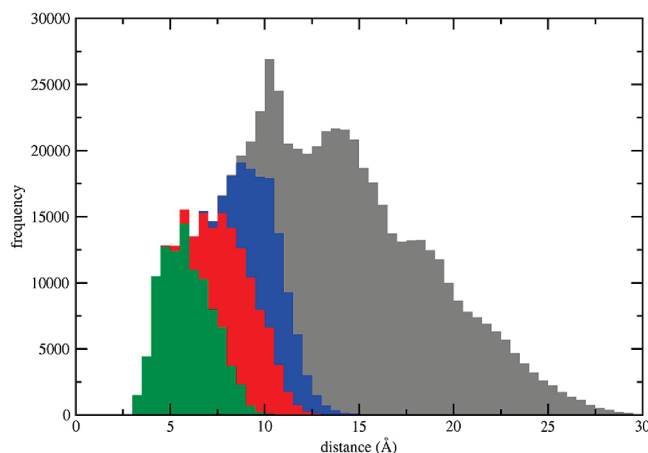
Figure 4. Pair distance distribution histograms for the representative side-chain atom model (see text). Histograms are sorted by the minimum GMD order of a representative atom pair (the smallest number k whose GMD includes the edge). Shown in color: minimum GMD order 2 (green), 3 (red), 4 (blue), and >4 (gray). The frequency values were sampled from trajectory 2 (peak distances and shapes of the pair distribution functions are trajectory invariant).

the time-dependent distance matrix and thereby determine the formation and breaking of graph edges. Figure 5A shows a separation distance time series of two representative atoms exhibiting typical level shifts. The two side chains form a contact from 600 to 1400 ns, but a direct assignment of contact formation and breaking using a cutoff of, for example, 8 Å would yield many spurious transitions within this time window due to the noise present on short time scales.

An abundance of alternative low-pass filtering and shift-detection methods have been proposed.[22,23] We tested two well-known and efficient filters for smoothing the time series, the moving average and the median, both defined within a sliding window. In this work, the median is defined as the smallest number in a series such that at least half the numbers are no greater than it. The median is influenced only by the ranking in the sample, making it robust against outliers. The moving average, on the other hand, is a linear filter, and thus it can be easily parallelized if desired. Figure 5B shows the performance of the moving average and median filter as applied to the level shift near 600 ns, indicating that the nonlinear median filter offers a satisfactory preservation of the shift.

In the following section, we implement the median filter for suppressing high-frequency noise in the distance matrix time series. The window half-width, δ, is an important time scale parameter defined by the user which controls the number of events that are detected. In preliminary testing, we have found that half-widths on the order of $10-100$ ns provide a reduction of spurious transitions by $2-3$ orders of magnitude (Supporting Information Figure 1) relative to the absence of a filter. The choice of δ depends on the time scale of the molecular process investigated by the user.

2.6. Suppressing Trivial Recrossings. One of the well-known problems in transition-state theory[24] is the overcounting of spurious recrossings at the boundary between two states.[12,25] Such recrossings may occur even after median

filtering, e.g., in the case of cutoff-based contact graphs when the cutoff is close to the mean of a distance distribution. The "event log" file (see Supporting Information) gives an example of repeated formation and breaking of the same contacts in the absence of any suppression of such recrossings. An overcounting of transitions occurs also for GMD-based events, since the Delaunay tetrahedralization is sensitive to representative atom motions. Several approaches have been proposed to remedy this problem, including the "almost Delaunay" triangulation by Bandyopadhyay and Snoeyink.[26] Here we take a different approach, exploiting the time dependence of the underlying model.

A large number of recrossings is simply an indication that a classification into contacts and noncontacts is not sufficient for the intended purpose of tracking "significant" level shifts. To compensate for these unwanted effects, we have developed a "trivial recrossing suppression" scheme (see Supporting Information). The idea, discussed in the "stable states picture" of chemical reactions[25] and recently used in the construction of Markov models from MD simulations,[12] is to introduce a buffer region and to track crossings until this buffer has been crossed completely. Figure 6 provides an overview of the nine possible paths crossing the buffer and identifies the remaining "nontrivial" contact formation and breaking events (green and red arrows, respectively), after application of the recrossing filter. Numeric labels assigned to the regions by our algorithm (Supporting Information) are also shown.

The use of a buffer requires the definition of a "contact metric" that separates the buffer from contacts and noncontacts. The metric may be continuous, as in the case of cutoffs, or discrete, as in the case of GMD graphs, where we use the minimum GMD order of an edge as metric (the smallest number k whose GMD includes the edge). The width of the buffer region is a free parameter defined by the user. In tests using cutoff-based contacts and the stationary trajectory 2 (which exhibits little activity and is thus a good test system for detecting spurious recrossings), we have found that even very small buffer zones of $0.3-0.5$ Å are highly effective in eliminating unwanted recrossings (Supporting Information Figure 2). In the case of GMD, we found the smallest possible buffer with a minimum order 2 (contacts), 3 (buffer), and 4 or higher (noncontacts) to be effective; it will be denoted as the "$k = 3$" crossing buffer in the following discussion.

2.7. Kernel-Based Activity Measure. The analysis described so far yields a detailed listing of K broken and formed contacts at corresponding times t_i ($i = 1, ..., K$). The cutoff- or GMD-based activities (rates of events) are computed from the event times by smoothing with a Gaussian kernel:

$$a(t) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{K} e^{-(t - t_i)^2/2\sigma^2} \tag{1}$$

The activity $a(t)$ is not normalized to unity as in probability density estimation, but to K, the total number of events, such that a gives the number of events per frame. The kernel standard deviation σ is matched to the median half-width δ as follows. The median filter can be considered a low-pass
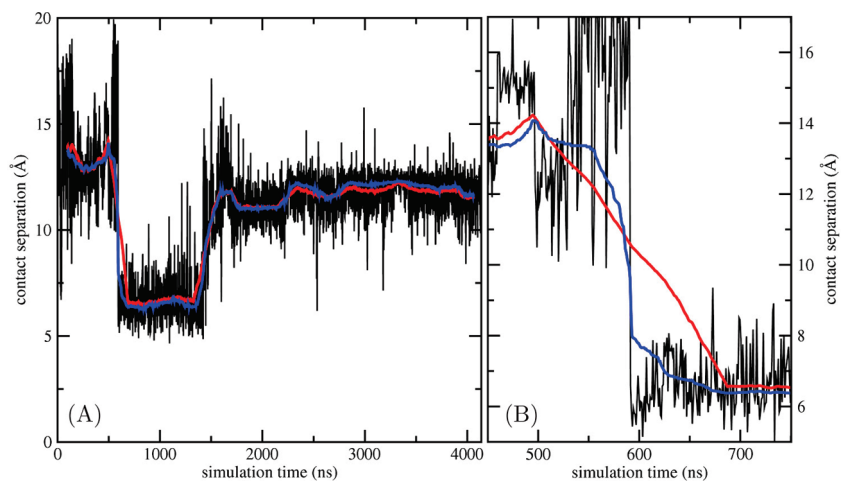
Figure 5. Smoothing of a typical contact time series (black) by moving average (red) and median filters (blue): (A) Full time window; (B) detailed view of a level shift at 600 ns. Shown is the separation of model atoms representing Asp5 and Phe10 in trajectory 3. The moving average and median filters used a sliding window of half-width = 100 ns.
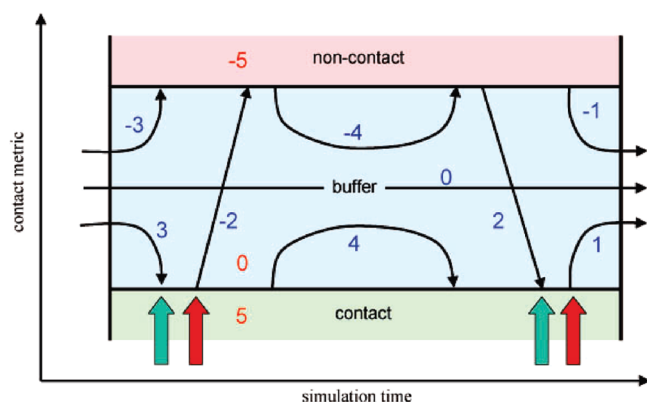

Figure 6. Suppression of trivial recrossings using a buffer zone (blue) between contact (green) and noncontact (red) zones. Nine types of buffer boundary crossings (thin black arrows) are theoretically possible. The colored arrows mark the time of four designated crossings (green, contact formation; red, contact breaking). All other crossings are suppressed (see text). Red numbers show initial numeric values used in the bidirectional tracking (see Supporting Information). Blue numbers are the final labels assigned to each of the nine crossing types. The metric for assigning zone boundaries may be continuous (distance cutoff) or discrete (GMD minimum order).

filter that attenuates frequencies above $(2\delta)^{-1}$ (the inverse of the median window width). The minimum sampling rate (or Nyquist rate) should be twice this frequency, or δ^{-1}, according to the Nyquist−Shannon sampling theorem. The "full width at half-maximum" (fwhm) parameter is commonly used to describe the resolving width of a kernel. This width must be small enough to resolve Nyquist rate samples. For the Gaussian kernel, the fwhm = $2(2 \ln 2)^{1/2}\sigma$ is thus matched to the inverse Nyquist rate: fwhm = δ.

Although it would, in principle, be possible to sample above the Nyquist rate (i.e., δ could be considered an upper bound for the kernel fwhm), we note that the smoothness of the activity curves is critical for estimating basin minima and basin transitions corresponding to local extrema of $a(t)$, so the maximum fwhm = δ is chosen in our application to

ensure the maximum smoothness of $a(t)$ (see Results and Discussion). The smoothing parameter δ thus corresponds to both the half-width of a median filter and the fwhm of a Gaussian kernel in our application.

2.8. Output. Our implementation provides a number of output files for inspection, plotting, and visualization of the methods described above: (a) a detailed log file of formation and breaking of contacts (for an example, see Supporting Information); (b) an activity time series data file containing the frame number, combined activity $a(t)$, and separate activities derived from either formation or breaking events; (c) trajectory files containing basin minima and basin transitions corresponding to local extrema of the combined activity $a(t)$; (d) a VMD-readable[27] contact graph for each frame (Figure 3A), enabling animation of contact graphs.

In the following section, we illustrate the use of the proposed analysis tools in practical MD applications.

3. Results and Discussion

The major idea associated with the tools introduced in the previous section is their ability to decompose the overall dynamics (expressed by the activity curves $a(t)$ of eq 1) into constituent individual events related to the breaking and formation of amino acid side-chain contacts. Before assessing the utility of detailed event logs in the practical analysis workflow, it is useful to compare the activities $a(t)$ to more traditional rms alignment techniques. Any similarities with the traditional techniques are nontrivial due to the different methodological paths taken by our methods. Differences, on the other hand, will suggest application areas for which our strategies are uniquely specified. We will describe two especially advantageous applications, the visualization of activity measures and the identification of activity basins and transitions in the trajectory.

3.1. Comparison of Tools for Activity Analysis. Figures 7 and 8 show the results of GMD-based (A) and cutoff-based (B) activity analysis applied to the "generic" trajectory 1 and the "folding" trajectory 4 (results for trajectories 2 and 3 are shown in Supporting Information Figure 2 and in Figure 9, respectively). For comparison with traditional
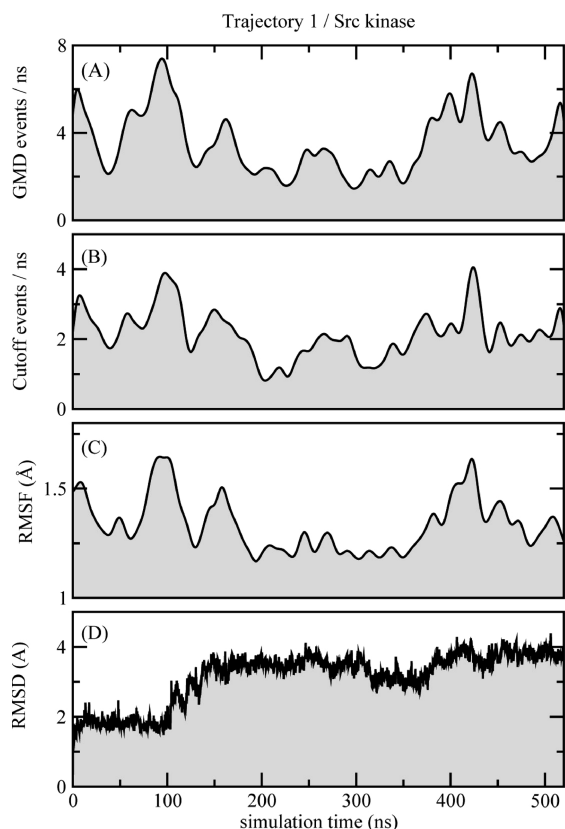
Event Detection in MD Simulations

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2601**

Figure 7. Comparison of conformational analysis tools applied to trajectory 1: (A) GMD-based activity ($k = 3$ crossing buffer); (B) cutoff-based activity (6.0−7.0 Å crossing buffer); (C) rms fluctuation in a sliding window; (D) α-carbon rms deviation from PDB entry 1Y57. The smoothing parameter δ setting Gaussian fwhm and median half-widths (see text) was 12.5 ns.

Figure 8. Comparison of conformational analysis tools applied to trajectory 4: (A) GMD-based activity ($k = 3$ crossing buffer); (B) cutoff-based activity (7.5−8.5 Å crossing buffer); (C) rms fluctuation in a sliding window; (D) α-carbon rms deviation from PDB entry 2F4K (with "A" variants of dual occupancy rotamers). The smoothing parameter δ setting Gaussian fwhm and median half-widths (see text) was 100 ns.

techniques, Figures 7 and 8 present also the rms fluctuation (C) and rms deviation from the native structure (D). The rms fluctuation in C measures the all-atom variability of consecutive frames in the trajectory weighted by a sliding Gaussian envelope function. To provide comparable detail, we have again matched the fwhm of the Gaussian envelope to the smoothing parameter δ (see above). We note that the unusual choice of a Gaussian envelope function for smoothing the rms fluctuations is critical for allowing comparison between these curves. If we used a more traditional sliding box envelope for the rms fluctuations, the curves in C would exhibit high-frequency noise (not shown), reducing the similarity with those in A and B.

The "generic" Src kinase trajectory 1 in Figure 7 represents a frequently encountered MD scenario and is thus of particular utility for the comparison of analysis tools. We describe similarities of analysis techniques by the Pearson correlation coefficient. The GMD-based (A) and cutoff-based activities (B) are quite similar in this case (correlation 0.86). Likewise, both activities are similar to the time-dependent rms fluctuation (C; correlations 0.90 and 0.84 for GMD- and cutoff-based activity, respectively). It is reassuring that the three measures (Figure 7A−C) are consistent in their characterization of traditional MD trajectories, even though there are considerable methodological differences in their

design. A minor difference from the two activity measures is the elevated background level exhibited by the rms fluctuation (Figure 7C), but this is inconsequential for analysis. Differences are more pronounced when comparing the three measures (Figure 7A−C) to the rms deviation. The first three measures show increased activity preceding a pronounced conformational change evident in the rms deviation (Figure 7D) after 100 ns (see also below). The subsequent activity peaks are not seen to have any major effect on the rms deviation. For example, the peak at 420 ns is due to local fluctuations in the disordered C-terminus which do not affect the rms deviation, since the structure has already moved far from the native conformation at this point.

The villin folding trajectory 4 in Figure 8 offers an opportunity to analyze a trajectory going from an extended to a compact, native state. The rms deviation (D) shows that the protein folds at 1400 ns. The GMD-based (A) and cutoff-based (B) activities yield a more detailed picture of the dynamic activity of the system up to 1400 ns, although the measures exhibit striking differences in this case (correlation 0.64). The major difference at the beginning of the trajectory is due to the fact that most contacts are outside the cutoff range in the initial extended conformation, but such folding events are included in the GMD, which does not depend on

Figure 9. Activity levels exhibited by the diffusive trajectory 3 during the first 1.25 μs: total GMD activity (gray); cutoff-based contact formation activity (green); cutoff-based contact breaking activity (red, plotted in negative direction to simplify comparison). A median filter (see text) was appl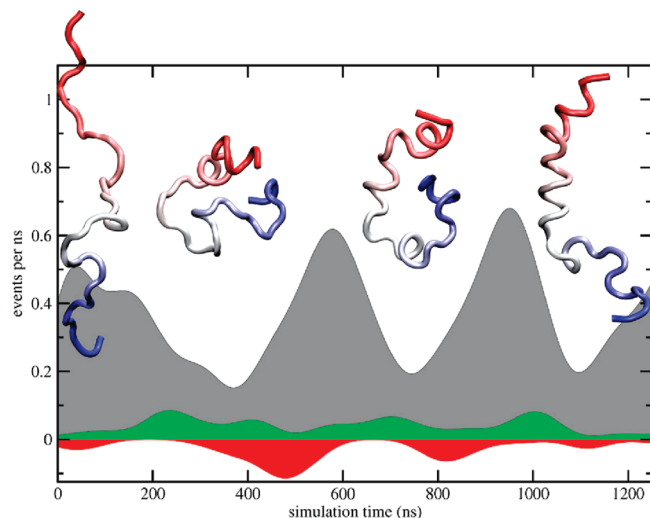ied, using a half-width of 100 ns. Recrossing suppression used a buffer of 6.0−7.0 Å (cutoff) or in the case of GMD, a buffer of $k = 3$ (see text). Snapshots of the trajectory above the plot correspond to the initial conformation and to three local minima of GMD activity that represent the basins directly below them. Molecular graphics renderings were created with VMD.[27] An animated AVI version of this figure, showing the full length of the trajectory, is available in the Supporting Information.

the cutoff. This difference highlights the adaptive property of contacts employed by the GMD approach. The rms fluctuation (C) is more similar (correlation 0.91) to the GMD- than to the cutoff-based activity (correlation 0.46). Despite the relatively high correlation, the local variations of the rms fluctuation (C) are significantly attenuated in this example compared to the variations of the GMD activity (A), in contrast to Figure 7, where the two measures show similar variability. Also, a small activity peak at 2200 ns is missed by the rms fluctuation. As described in subsection 3.2, this activity peak corresponds to substantial fluctuations of the first helix.

Our results suggest that the rms deviation is the least reliable predictor of conformational transitions because it misses some events detected by the other measures once the rms deviation reaches high numeric values. Also, GMD- and cutoff-based activities provide some additional information one could not obtain from rms fluctuations. The GMD- and cutoff-based activities differ especially in the folding trajectory 4. We explore differences between GMD- and cutoff-based activities further in subsection 3.3.

3.2. Utility of Detailed Event Logs. One important advantage of the proposed analysis is that it provides a detailed listing of constituent events that facilitates an underlying structural interpretation of the activity, beyond detection of periods of high activity itself. Traditional analysis tools based on Cartesian coordinates are not able to provide such detail. The cases of trajectories 1 and 4 illustrate the utility of event logs provided by the new algorithms. These can be particularly useful when combined

with expert knowledge, for example from mutagenesis data, of which residues are believed to play an important role.

The event logs of the Src kinase trajectory 1 indicate that Phe405 undergoes a conformational change that results in its exchange of packing partner during the 90−140 ns time period. Initially, Phe405 is in proximity of Glu310, Val313, Leu317, and Met314, contacts which are broken at 89, 90, 96, and 136 ns, respectively. The loss of contacts is compensated by the formation of a new contact with His384 at 96 ns simulation time. This conformational change mainly involving Phe405, His384, and Met314 is highly intriguing and potentially important, since Phe405 and His384 belong to the well-known DFG and HRD motifs that are almost universally conserved among protein kinases, and Met314, Phe405, and His384 are all part of a critical structural "spine" that was identified to stabilize kinase active structures.[28]

For the villin folding trajectory 4, formation of helical ($i,i+3$) and ($i,i+4$) contacts contribute substantially to the activity in the initial part of the trajectory. This is not directly followed by folding, but rather the protein appears temporarily trapped due to the formation of nonnative interactions. Specifically, after approximately 600 ns, a contact forms between the oppositely charged N- and C-terminal residues. This contact, together with an overextension of helix 1 through to residue Thr13, characterizes a persistent nonnative state between approximately 900 and 1100 ns that is associated with a dip in activity (Figure 8). Exit from this state is accompanied by the loss of the nonnative helical contacts in helix 1 and the subsequent formation of helix 2. The final event in folding is the unraveling and re-formation of helix 1, together with a reorientation of the loop between helices 1 and 2. This is accompanied by a burst of contact formation between hydrophobic residues, including the Phe6-Phe17 contact in the core, which is formed at ∼1400 ns. After folding, helix 1 occasionally undergoes substantial fluctuations, leading to the rise in activity at approximately 2200 ns visible in Figure 8. This involves the partial transient loss of helical structure from helix 1, reflected in changes in the contacts in that helix, accompanied by a change in orientation of helix 1 with respect to the rest of the protein that is reflected in changing contacts between residues at the beginning of helix 1 with those near the beginning of helix 2.

Once the contact formation or breaking events are identified, geometric inspection tools such as those provided by VMD[27] may add to the interpretation. It would have been impossible to extract this highly specific information with one of the traditional rms deviation or rms fluctuation measures.

3.3. Visualization of Activity Results. Given the differences between GMD and cutoff when applied to folding trajectory 4 (Figure 8B), we have investigated the discrepancy further using the "diffusive" trajectory 3. Since the original level shifts that give rise to activities can be separated into formation and breaking events, we considered separately the formation and breaking activities derived from the two classes. The differences were striking for cutoff-based activity levels (Supporting Information Figure 3; correlation 0.07 between formation and breaking), whereas in the case of

GMD the formation and breaking contributions were very similar (Supporting Information Figure 4; correlation 0.77). This indicates that, at least in the case of folding trajectories, the formation and breaking of cutoff-based contacts are asymmetric and at times either one may be dominating, whereas in the GMD graph the total number of contacts is nearly constant. We thus propose to visualize separate formation and breaking activities in the case of cutoff contacts, and only the total activity of the GMD.

Figure 9 displays such a "combination plot" of activities together with snapshots of the trajectory at low GMD activity. An animated AVI version of this visualization is available in the Supporting Information. One can observe at several times the pronounced asymmetry in the cutoff activity levels. A dominant formation of cutoff contacts, such as at 250, 700, and 1000 ns, typically precipitates a stabilization of the system (as judged by low GMD activity at 350, 750, and 1100 ns). Likewise, a dominant breaking of cutoff contacts, such as at 500 and 800 ns, clearly favors subsequent folding transitions (corresponding to high GMD activity at 600 and 900 ns). The proposed combination plot thus provides a nuanced characterization of folding activity, in which periods of stabilization or destabilization of the overall fold can be matched with more detailed changes in the side-chain packing.

The results suggest that inactive periods observed in folding trajectories are caused by preceding periods of contact formation of the structure, whereas large-scale folding transitions follow after periods of contact destabilization. The observed dependence of structural stability on contact formation could be used to enhance sampling in folding trajectories.

3.4. Segmentation of Activity Basins and Transitions. Figure 9 suggests a natural segmentation of the trajectory into quiescent "basins" separated by "transitions." We assigned local minima (basin centers) and local maxima (transitions) using a finite difference approximation of the first derivative of the total (GMD- or cutoff-based) activity $a(t)$. The local maxima correspond to highly active periods of the trajectory that separate basins of inactivity. The local minima roughly correspond to the structures with the greatest contact similarity to the average structure of the local basin. These minima are shown in Figure 9 above the GMD activity plot, representing the inactive basins directly below them. This strategy can also be applied (after Gaussian smoothing) to the traditional rms fluctuation.

For typical MD trajectories such as trajectory 1, the maxima and minima of $a(t)$ are not very sensitive to the graph method used. For example, 75% of the minima and transitions derived from the GMD activity (Figure 7A) can be found to be within 5 ns simulation time of like extrema exhibited by the rms fluctuation (Figure 7C). The similarity with the rms fluctuation was somewhat less pronounced for the cutoff-based activity (63%; Figure 7B). As can be expected, the observed conservation of minima and transitions agrees qualitatively with the above Pearson correlation analysis. We propose to use the GMD activities for the assignment of basins and transitions whenever possible, due

to the more pronounced undulations relative to the rms fluctuation (Figure 8A,C).

4. Conclusion

We have introduced tools for automated event detection and activity monitoring in MD simulations and demonstrated their application to state-of-the-art trajectories. Our method introduces intuitive parameters to be defined by the user, as follows:

(a) The type of contact graph. We recommend a cutoff-based graph to detect detailed side-chain contact formation and breaking, or a GMD-based graph to detect global activity.

(b) The designated crossing buffer. We recommend 6−7 Å cutoffs or GMD order $k = 3$.

(c) The temporal smoothing parameter δ. This value depends on the length of the simulation and the desired level of detail.

(d) The side-chain atom selection. We provide a default profile for standard amino acid residues, which may be modified for specific systems or nonstandard residues.

All other steps in the methodology are automated, including median filtering, suppression of trivial recrossings, kernel activity estimation, calculation of basin minima and transitions, and data file output.

Our current serial implementation is sufficiently efficient to allow for the analysis of microsecond-scale trajectories. An analysis of 4496 frames of trajectory 3 took only a few minutes of compute time on a standard Linux workstation. For much longer trajectories, we expect that parallelization of the analysis may be required; such parallelization should be straightforward using, for example, the recently developed HiMach framework.[29]

Our implementation brings together state-of-the-art methodologies from time-series analysis, computational geometry, graph theory, and biochemistry to address the activity-monitoring and event-detection problem. The limitations of our methods include the focus on global rearrangements in the structure; some events of scientific interest leave only very small footprints in the surrounding protein matrix. Ion and solvent diffusion through membrane channels, for example, would require different detection techniques. In addition, the parameters of our method have not yet been optimized for lipids and nucleic acids, although it would in principle be possible to generalize the coarse-grained model to nonprotein contacts—especially in the case of GMD, which is independent of specific cutoff distances.

An additional limitation of our analysis is that events are still relatively frequent for human interpretation (about 100−1000 events were observed per microsecond). For longer trajectories, it may be helpful to further reduce the complexity of the contact patterns using one or more of the following strategies: (i) ignoring contacts formed by residues with nonexistent or short side chains such as Cys, Pro, and Ala; (ii) substantially increasing the crossing buffer; (iii) ranking events by the sequence conservation of participating residues, the energy levels of participating residues, or a correlation analysis of the motion of participating residues.

Our tests on four trajectories have revealed a number of advantages of our activity-based calculations relative to the

more traditional rms fluctuation. These include (i) a higher sensitivity at low activity levels (Figure 8A,C); (ii) a reduced background noise contribution (Figure 7A, C); (iii) a detailed listing of individual events underlying the observed activity; (iv) coarse model calculations that are roughly an order of magnitude faster than an all-atom analysis; and (v) a functionally relevant diversification of the tool arsenal: the GMD activities show an overall fold rearrangement, the cutoff activities measure contact formation and breaking, and the traditional rms fluctuation measures the variability of Cartesian coordinates of neighboring frames. The importance of automated analysis techniques will only grow as efforts in high-throughput MD simulation—such as the "Dynameom- ics" project[30]—make large numbers of MD trajectories publicly available for mining and interrogation.

4.1. Dissemination. All tools described in this article will be documented and freely distributed as part of the Python-based "TimeScapes" package at URL http:// www.DEShawResearch.com (Resources). TimeScapes is capable of reading the trajectories produced by many popular MD programs, including AMBER, CHARMM, NAMD, X-PLOR, Desmond, LAMMPS, and GROMACS, making the package widely applicable.

References

(1) Karplus, M.; McCammon, J. A. Molecular Dynamics Simula- tions of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(2) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(3) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. The Theory and Practice of Distance Geometry. *Bull. Math. Biol.* **1983**, *45*, 665–720.

(4) Holm, L.; Sander, C. Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **1993**, *233*, 123–138.

(5) Keller, P. A.; Leach, S. P.; Luu, T. T.; Titmuss, S. J.; Griffith, R. Development of Computational and Graphical Tools for Analysis of Movement and Flexibility in Large Molecules. *J. Mol. Graphics Modell.* **2000**, *18*, 235–241.

(6) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-microsecond Protein Folding. *J. Mol. Biol.* **2006**, *359*, 546–553.

(7) Singh, J.; Thornton, J. M. *Atlas of Protein Side-Chain Interactions*, Vols. I and II; IRL Press: Oxford, U.K., 1992.

(8) Yang, H.; Parthasarathy, S. Mining Spatial and Spatio- Temporal Patterns in Scientific Data. *Proceedings of the 22nd International Conference on Data Engineering Workshops*, Atlanta, GA; IEEE Computer Society: Washington, D.C., 2006.

(9) Zhou, R.; Parida, L.; Kapila, K.; Mudur, S. PROTERAN Animated Terrain Evolution for Visual Analysis of Patterns

(10) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146–168.

(11) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identi- fication of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains. *Lin. Alg. Appl.* **2000**, *315*, 39– 59.

(12) Buchete, N.-V.; Hummer, G. Peptide Folding Kinetics from Replica Exchange Molecular Dynamics. *Phys. Rev. E* **2008**, *77*, 030902(R).

(13) Yaliraki, S. N.; Barahona, M. Chemistry Across Scales: From Molecules to Cells. *Philos. Trans. R. Soc. A* **2007**, *365*, 2921– 2934.

(14) Factor, A. D.; Mehler, E. L. Graphical Representation of Hydrogen Bonding Patterns in Proteins. *Protein Eng.* **1991**, *4*, 421–425.

(15) Sessions, R. B.; Gibbs, N.; Dempsey, C. E. Hydrogen Bonding in Helical Polypeptides from Molecular Dynamics Simulations and Amide Hydrogen Exchange Analysis: Alamethicin and Melittin in Methanol. *Biophys. J.* **1998**, *74*, 138–152.

(16) Kaźumierkiewicz, R.; Liwo, A.; Scheraga, H. A. Addition of Side Chains to a Known Backbone with Defined Side-Chain Centroids. *Biophys. Chem.* **2003**, *100*, 261–280.

(17) Jeong, J. I.; Jang, Y.; Kim, M. K. A Connection Rule for Alpha-Carbon Coarse-Grained Elastic Network Models Using Chemical Bond Information. *J. Mol. Graphics Modell.* **2006**, *24*, 296–306.

(18) Martinetz, T.; Schulten, K. Topology Representing Networks. *Neural Networks* **1994**, *7*, 507–522.

(19) Fischer, I.; Gotsman, C. Fast Approximation of High Order Voronoi Diagrams and Distance Transforms on the GPU. *J. Graphics Tools* **2006**, *11*, 39–60.

(20) Martinetz, T. Competitive Hebbian Learning Rule Forms Perfectly Topology Preserving Maps. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN-93)*; Gielen, S., Kappen, B., Eds.; Springer Verlag: Heidelberg, Germany, 1993; pp 427−434.

(21) de Berg, M.; van Kreveld, M.; Overmars, M.; Schwarzkopf, O. *Computational Geometry: Algorithms and Applications*; Springer Verlag: Berlin, 2000.

(22) Gather, U.; Fried, R.; Lanius, V. Robust Detail-Preserving Signal Extraction. In *Handbook of Time Series Analysis*; Schelter, B., Winterhalder, M., Timmer, J., Eds.; Wiley-VCH: Weinheim, Germany, 2006; pp 131−157.

(23) Ye, L.; Wu, Z.; Eleftheriou, M.; Zhou, R. Single-Mutation- Induced Stability Loss in Protein Lysozyme. *Biochem. Soc. Trans.* **2007**, *35*, 1551–1557.

(24) Pollak, E.; Talkner, P. Reaction Rate Theory: What It Was, Where Is It Today, and Where Is It Going. *Chaos* **2005**, *15*, 026116.

(25) Northrup, S. H.; Hynes, J. T. The Stable States Picture of Chemical Reactions. I. Formulation for Rate Constants and Initial Condition Effects. *J. Chem. Phys.* **1980**, *73*, 2700– 2714.

(26) Bandyopadhyay, D.; Snoeyink, J. Almost-Delaunay Simplices: Nearest Neighbor Relations for Imprecise Points. *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete*

Event Detection in MD Simulations

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2605**

Algorithms, New Orleans, LA; Society for Industrial and Applied Mathematics: Philadelphia, 2004; pp 410−419.

(27) Humphrey, W. F.; Dalke, A.; Schulten, K. VMD-Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(28) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Ten Eyck, L. F. Surface Comparison of Active and Inactive Protein Kinases Identifies a Conserved Activation Mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17783–17788.

(29) Tu, T.; Rendleman, C. A.; Borhani, D. W.; Dror, R. O.; Gullingsrud, J.; Jensen, M. O.; Klepeis, J. L.; Maragakis, P.; Miller, P.; Stafford, K. A.; Shaw, D. E. A Scalable Parallel Framework for Analyzing Terascale Molecular Dynamics Trajectories. *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, Austin, TX; ACM Press: New York, NY, 2008.

(30) Beck, D. A. C.; Jonsson, A. L.; Schaeffer, R. D.; Scott, K. A.; Day, R.; Toofanny, R. D.; Alonso, D. O. V.; Daggett, V. Dynameomics: Mass Annotation of Protein Dynamics and Unfolding in Water by High-Throughput Atomistic Molecular Dynamics Simulations. *Protein Eng. Des. Sel.* **2008**, *21*, 353–368.

JCTC Journal of Chemical Theory and Computation

SIMUFLEX: Algorithms and Tools for Simulation of the Conformation and Dynamics of Flexible Molecules and Nanoparticles in Dilute Solution

José García de la Torre,*,[||] José G. Hernández Cifre,[||] Álvaro Ortega,[||]
Ricardo Rodríguez Schmidt,[||] Miguel X. Fernandes,[†] Horacio E. Pérez Sánchez,[‡] and
R. Pamies[§]

*Departamento de Química Física, Facultad de Química Universidad de Murcia,
30071 Murcia, Spain, Centro de Química da Madeira, Universidade da Madeira,
9000-390 Funchal, Portugal, Forschungszentrum Karlsruhe GmbH, Institut für
Nanotechnologie, D-76021 Karlsruhe, Germany, and Department of Physical
Chemistry, University of Oslo, Oslo, Norway*

Abstract: A computer programs suite, SIMUFLEX, has been constructed for the calculation of solution properties of flexible macromolecules modeled as bead-and-connector models of arbitrary topology. The suite consists mainly of two independent programs, BROWFLEX that generates the macromolecular trajectory by using the Brownian dynamics technique and ANAFLEX that analyzes that trajectory to get solution properties of the macromolecule. In this paper, we describe theoretical aspects about the macromolecular model and the Brownian dynamics algorithm used and describe some of the numerous properties that can be evaluated. In order to provide examples of the application of the methodology, we present simulations of dynamic properties of DNA with length ranging from 10 to 10^5 base pairs. SIMUFLEX is able to run simulations with more or less coarse-grained models, thus enabling such multiple-scale studies.

1. Introduction

Solution properties of macromolecules (hydrodynamic coefficients, intrinsic viscosity, radiation scattering-related quantities...) shed light about their shape and conformation[1] and therefore are a primary source of information to predict their solution behavior. For example, the determination of such properties are of fundamental relevance when treating with biological macromolecules (DNA, proteins...), since their physiological functions are closely related to the solution conformations that they can adopt.[2] A powerful tool that helps to predict and understand macromolecular structure and dynamics is computational modeling and simulation.[3]

There exist well developed procedures based on bead modeling[4−8] or alternative approaches[9−12] to predict hydrodynamic properties of rigid macromolecules and nanoparticles. However, most synthetic polymers and many biological macromolecules are flexible and do not present a defined shape. Therefore, the development of computational procedures to predict the solution behavior of flexible and semiflexible macromolecules is of great interest. The large size typical of macromolecules and nanoparticles, and the long times typical of their dynamics, precludes usually the use of atomic-level models, and the conformational variability of flexible entities adds further complexity. Thus, the prediction of solution properties requires simplified schemes, based on more or less coarsely grained models. The classical bead-and-spring model of polymer physics, in which the model elements represents large pieces (subchains) of the long polymer chain,[13,14] is a very coarse grained model. Nowadays, the coarse-grained modeling concept is being

* Corresponding author e-mail: jgt@um.es.
[||] Facultad de Química de Universidad de Murcia.
[†] Universidade da Madeira.
[‡] Forschungszentrum Karlsruhe GmbH, Institut für Nanotechnologie.
[§] University of Oslo.

applied with more detail, with models whose elements represent, for instance, the repeating units - amino acid or nucleotide residues - of biomacromolecules.[15] On the other hand, useful schemes to build coarse-grained models have been recently developed.[16,17] Because of the widespread utilization - over the past two decades - of atomistic molecular dynamics simulations, there are many commercial and public domain tools for that purpose. However, for multiscale, coarse-graining simulation, one misses a wide availability of similar tools. Thus, we have intended[18] to develop computational methodologies where flexible macromolecules are represented at such a coarse-grained level as bead-and-connector models and to predict their solution behavior by simulation techniques as Monte Carlo (MC) and Brownian dynamics (BD).[18]

When the simulation of flexible entities is restricted to the prediction of conformational, equilibrium properties, and some overall hydrodynamic coefficients, Monte Carlo methods are applicable. In order to provide a tool for the MC simulation of quite general flexible bead-and-connector models, we recently published the public-domain program MONTEHYDRO,[19] that implements an importance sampling Monte Carlo procedure for the generation of random conformations of flexible structures, which includes the calculation of overall hydrodynamic properties in the so-called rigid-body treatment,[20−22] obtained as conformational averages over the values calculated for instantaneous conformations considered as rigid structures.[5,23]

However, to study dynamic aspects of flexible macromolecules in solution, such as relaxation processes and non-equilibrium behavior, it is necessary to solve the equation of motion that governs the macromolecular dynamics. This can be done by using molecular dynamics (MD) or Brownian dynamics (BD).[24] Because of the above-mentioned drawbacks, MD is not adequate for long time and size scales. BD is a numerical technique to solve the stochastic equation of motion that arises from considering the solvent as a continuum, thus eliminating the solvent degrees of freedom which allows for reaching longer times in the simulated physical system. In other words, BD simulations describe the Brownian motion of a collective of frictional elements, beads in our model, which can interact with each other through different potentials.

An essential aspect in the BD simulation is the inclusion of the so-called hydrodynamic interaction (HI) effect, which determines the solvent-mediated influence of the motion of every element of the model on the others. Our group[25−29] was among others[30−34] who pioneered the use of BD simulations including hydrodynamic interaction (HI) effects to predict dynamic properties of macromolecules in solution. As it is known from polymer hydrodynamic theory,[14,35] and confirmed by BD simulations,[28,36,37] the rigorous inclusion of the HI effect (avoiding approximations, like that of conformational preaverage) is essential for the accurate prediction of hydrodynamic properties results comparable to experiments. Nevertheless, BD simulations without inclusion of HI sample correctly the configurational space, so that some authors have proposed that BD could be used as a smart Monte Carlo method.[38] This adds a further utility to BD

methodologies, providing an efficient way to obtain also equilibrium conformational properties.

Along our previous works we have been developing a BD simulation scheme that enables for the calculation of solution properties of flexible macromolecules with arbitrary complexity. Our procedures take into account fluctuating (nonpreaveraged) hydrodynamic interaction as well as the possibility of including different types of intramolecular potentials to represent excluded volume conditions (solvent quality) and electrostatic interactions. That computational scheme is implemented in a suite of public domain (freely available from our Web server, see below), named SIMU-FLEX, which is presented in this paper. The suite consists mainly of two programs BROWFLEX and ANAFLEX. The program BROWFLEX generates a Brownian trajectory of a flexible bead-and-connector model with arbitrary connectivity, and the program ANAFLEX analyzes that trajectory to obtain several steady and time-dependent macromolecular quantities. In this way, many conformational and hydrodynamic solution properties, from single-valued coefficients to more complex experiments as well as different time correlation functions, can be straightforwardly evaluated from the Brownian trajectory. Furthermore, a most interesting feature of the BD technique is that it allows the simulation of the behavior of an individual molecule,[39,40] which is of great importance due to the emergence of single-molecule experimental techniques.[41] At this respect, SIMUFLEX is an useful tool to study single-molecule behavior of flexible macromolecules with arbitrary topology. Thus, our contribution joins those of other groups who have published Brownian dynamics simulation programs with different scope or structure[42,43] and focused on particular macromolecular systems (for instance the UHBD package[42] is appropriate for studying protein−protein association). On the other hand, the SIMUFLEX package was devised to treat with a variety of macromolecular models and physical situations, for example the presence of external agents, as well as to analyze easily an amount of macromolecular properties including a number of commonly employed correlation functions.

In this paper, we first describe some theoretical aspects of the modeling and BD simulation methodology implemented in SIMUFLEX. Then, we present several examples, all concerning the dynamics of DNA molecules in solution. In order to illustrate the multiscale possibilities of SIMU-FLEX, the examples span a wide range of DNA sizes and cover both bulk-solution and single-molecule properties.

2. Models and Simulation Methods

In this section we specify the two main features in the simulated model. The first one corresponds to the mechanical or energetic features pertaining to the molecule itself and, eventually, its interaction with external agents (e.g., fields, walls, etc.). The second group of aspects comprise those relative to the motion of the molecular model in the viscous solvent, like viscous drag, hydrodynamic interaction, Brownian motion, etc., which are key factors for the construction of the simulation algorithm.

2.1. Mechanical Model: The Force Field. The simulation model is composed by what we generically call elements,
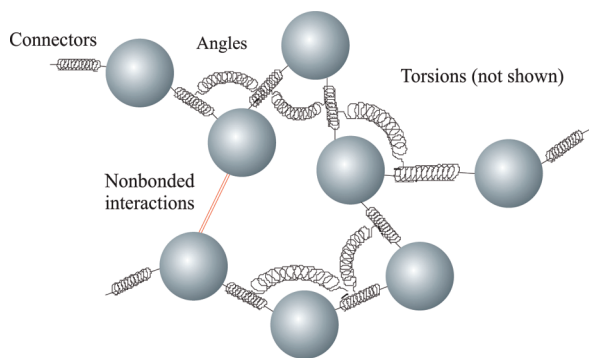
Figure 1. Pictorial representation of the generalized bead-and-spring model.

which will be later considered as spheres, or "beads", for the description of their hydrodynamic behavior. The N elements interact in a number of ways, which give rise to a potential energy that we decompose in several contributions

$$V^{(tot)} = \sum_{conn} V_{ij}^{(conn)} + \sum_{ang} V_{ijk}^{(ang)} + \sum_{tors} V_{ijkl}^{(tors)} + \sum_{EVpairs} V_{ij}^{(EVpair)} + \sum_{CHpairs} V_{ij}^{(CHpair)} + \sum_{i} V_{i}^{(E)} \quad (1)$$

The terms in eq 1 correspond to each kind of interaction, as described in the following paragraphs. A schematic overview of the model is displayed in Figure 1.

Primarily, the elements are joined by connectors, which describe the topology of the molecule or particle that is being modeled. A common case is that of linear chains, in which each element (except the terminal ones) is joined to its two neighbors. In general, an element may be joined to an arbitrary number of other elements. The sum extends over all the pairs of connected elements. Connectors behave mechanically as springs with an associated potential $V_{ij}^{(conn)}(l_{ij})$ that depends on the instantaneous distance between the two joined elements, equal to the length of the spring vector, $l_{ij} = |\mathbf{r}_{ij}| = |\mathbf{r}_j - \mathbf{r}_i|$, where $\mathbf{r}_i$ is the position vector of the i-th bead. Among others, BROWFLEX considers the following spring potential devised in a previous work to simulate dendrimer molecules[44]

$$V^{(conn)} = -\frac{1}{2}Hl_{max}^2 \ln\left(\frac{l_{max}^2 - l^2}{l_{max}^2 - l_e^2}\right) - \frac{1}{2}Hl_{max}l_e \ln\left[\frac{(l_{max} + l)(l_{max} - l_e)}{(l_{max} - l)(l_{max} + l_e)}\right] \quad (2)$$

In eq 2 subscripts ij, that should be attached to $V^{(conn)}$, l, l_e, l_{max}, and H, are omitted for the sake of legibility. The equilibrium length l_e ($V^{(conn)}(l_e) = 0$), the maximum elongation l_{max}, and the force constant H are the three parameters of this general spring potential, which we call "hard-FENE" because it includes, as particular cases, several commonly used spring types. When $l_{max} \rightarrow \infty$ (in practice, a sufficiently large number), it reduces to $V^{(conn)} = 1/2H(l - l_e)^2$, proper of a Hookean (Fraenkel)[45] spring that is usually employed, with a large value of H, to represent stiff connectors with an equilibrium length l_e (the rms fluctuation in spring length, $(\langle l^2 \rangle - \langle l \rangle^2)^{1/2} = H/(k_BT)$[25,46] is, for instance 10% of l_e for H

$= 100k_BT/l_e^2$). Furthermore, with $l_{max} \rightarrow \infty$ and $l_e = 0$ we have $V^{(conn)} = 1/2Hl^2$, which is the potential associated with a Gaussian distribution of the spring length, with $\langle l^2 \rangle = 3k_BT/H$ as used in the Rouse model[13] of linear polymer chains composed by Gaussian subchains. The Rouse model with a linear force and infinite extensibility is inappropriate when external agents, particularly strong flows, stretch the chain and the distribution is not Gaussian. For those cases, the most popular choice is the FENE (finitely extensible, nonlinear elastic; Warner)[47] spring, whose potential, $V^{(conn)} = -(1/2)Hl_{max}^2\ln(1 - l^2/l_{max}^2)$ is a particular case of eq 2 for $l_e = 0$. For the spring potential, as for the other pairwise potentials depending on the distance between elements, the forces acting on the two elements are $\mathbf{F}_i = -\mathbf{F}_j = [dV(r_{ij})/dr_{ij}]\mathbf{r}_{ij}/r_{ij}$, where the derivative of the potential adopts a long but immediate expression (omitted) that allows an easy calculation of the forces.

The angles, α_{ijk}, between two neighbor springs joining beads i and j, and j and k, may have an associated potential $V_{ijk}^{(ang)}(\alpha_{ijk})$. A simple and useful potential for the bending angle is the quadratic form, $V(\alpha) = (1/2)Q(\alpha - \alpha_0)^2$, where α_0 is the equilibrium value of the angle, and Q is a bending force constant. Again, the subscripts ijk are omitted for brevity, but we recall that there may be specific values of the parameters for each angle in the model. In order to make the program applicable to chemical entities (real molecules), we have also included in SIMUFLEX torsional potentials associated with hindered internal rotation. If ij, jk, and kl, are three consecutive bonds, internal rotation around the jk bond can be represented by a potential $V_{ijkl}^{(tors)}(\phi_{ijkl})$, where ϕ_{ijkl} is the internal rotation angle. In the program we have included a variety of $V(\phi)$ functions, corresponding to the most frequent kinds of chemical bonds. The expressions for the forces associated with bending and internal rotations can be found in the literature.[24,26]

The force field includes two other kinds of pairwise intramolecular potentials. One kind, denoted in eq 1 as $V_{ij}^{(EVpair)}$ is usually for excluded-volume (EV) interactions, for which BROWFLEX considers various possibilities, including the useful and meaningful Lennard-Jones potential

$$V^{(EVpair)} = 4\varepsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r}\right)^{12} - \left(\frac{\sigma_{LJ}}{r}\right)^{6}\right] \quad (3)$$

where ϵ_{LJ} and σ_{LJ} are the Lennard-Jones parameters, along with other forms, like hard-spheres, exponential or Gaussian repulsion, etc. The second kind of pairwise potentials indicated in eq 1 as $V_{ij}^{(CHpair)}$ is intended for any other interaction that would superimpose to the excluded volume effect. An example is the intramolecular electrostatic interaction between charged elements, which can be properly described by a screened Coulomb, Debye−Hückel potential

$$V^{(CHpair)} = (A/r)\exp(-\kappa r) \quad (4)$$

where A is related to the charge of the two elements and the dielectric constant of the solution, and κ, the screening parameter, is related to the ionic strength of the medium.[48] Again, in eqs 3 and 4 the subscripts ij are omitted, and each pair may have its own parameters.

SIMUFLEX: Algorithms for Flexible Molecules

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2609**

Finally, $V_i^{(E)}$ indicates any interaction between any individual element and an external agent (or constrain), since BD is an adequate technique to simulate macromolecules in, for example, electric fields[49,50] and in biomembranes.[51] BROWFLEX includes various useful possibilities like interaction of a charged element with an external electric field; walls that the element cannot trespass; and anchorage, by means of a hard spring, of an element to a fixed point. In the BROWFLEX user guide there is a detail relation of the forces included in the force field of BROWFLEX and how they must be used to build the chain model.

2.2. Hydrodynamics and Brownian Dynamics: The Algorithms. For the generation of Brownian trajectories of the mechanical molecular model in the viscous solvent, including hydrodynamic interaction effects, we propose the use of a procedure, based on the Ermak-McCammon[30,46] (E-M) algorithm, proposed by Iniesta and García de la Torre[29] (I-GT). In the E-M algorithm, the final position $\mathbf{r}_i$ of bead i after at time step Δt is calculated starting from its initial position $\mathbf{r}_i^0$, according to

$$\mathbf{r}_i = \mathbf{r}_i^0 + \frac{\Delta t}{k_B T}\sum_{j=1}^{N}\mathbf{D}_{ij}^0\cdot\mathbf{F}_j^0 + \Delta t\sum_{j=i}^{N}\left(\frac{\partial \mathbf{D}_{ij}}{\partial \mathbf{r}_j}\right)^0 + \mathbf{R}_i + (\Delta t)\mathbf{v}_i^0 \tag{5}$$

where $\mathbf{F}_j^0$ is the total force on bead i, $\mathbf{D}_{ij}$ is the 3×3 i,j-block of the $3N \times 3N$ grand diffusion tensor, $\mathbf{D}$, and $\mathbf{R}_i$ is a Gaussian random vector with zero mean and covariance

$$\langle \mathbf{R}_i\mathbf{R}_j\rangle = 2\mathbf{D}_{ij}\Delta t \quad i,j = 1, ..., N \tag{6}$$

If the dynamics takes place in a flowing solvent, we include in the E-M algorithm the term $\mathbf{v}_i^0$, which is the fluid velocity in the position of the bead i due to the flow field.

As the E-M is equivalent (without the Brownian drift term) to the first-order Euler algorithm for ordinary differential equations, Iniesta and García de la Torre proposed and algorithm inspired in the second-order Runge–Kutta procedure. In the I-GT procedure, each step is taken twice, in a predictor-corrector fashion. First, the predictor step is an E-M step, taking according to eq 5 that conducts to preliminary new bead positions $\mathbf{r}'_i$. Next, the forces, diffusion tensors, and their gradients are evaluated at these positions, and then the step is repeated, from the original initial position, taking the means of the quantities calculated before and after the predictor step (indicated with superscript '); thus, the second corrector step is given by

$$\mathbf{r}_i = \mathbf{r}_i^0 + \frac{\Delta t}{k_B T}\frac{1}{2}\sum_{j=1}^{N}(\mathbf{D}_{ij}^0\cdot\mathbf{F}_j^0 + \mathbf{D}'_{ij}\cdot\mathbf{F}'_j) +$$
$$\Delta t\frac{1}{2}\sum_{j=i}^{N}\left[\left(\frac{\partial \mathbf{D}_{ij}}{\partial \mathbf{r}_j}\right)^0 + \left(\frac{\partial \mathbf{D}_{ij}}{\partial \mathbf{r}_j}\right)'\right] + \mathbf{R}'_i + \Delta t\frac{1}{2}(\mathbf{v}_i^0 + \mathbf{v}'_i) \tag{7}$$

Although in the I-GT algorithm each step is taken twice, which amounts to duplicating the computing time per step, the time step Δt can be remarkably (over 1 order of magnitude) longer than in the first-order E-M procedure, so that the computing time needed to simulate a trajectory of a given duration is notably reduced. Several authors have commented on the advantages of the I-GT procedure.[52−55]

Fluctuating hydrodynamic interactions between beads can be accounted for by means of the Rotne-Prague-Yamakawa tensor,[56,57] valid when all elements or beads of the chain are equal size. García de la Torre and Bloomfield[58] extended that tensor to the case of nonequal elements. Using that HI tensor, the diffusion tensor that enters in the Brownian algorithm, $\mathbf{D}_{ij}$, reads

$$\mathbf{D}_{ij} = \frac{k_B T}{8\pi\eta_0 r_{ij}}\left[\mathbf{I} + \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}}{r_{ij}^2} + \frac{\sigma_i^2 + \sigma_j^2}{r_{ij}^2}\left(\frac{1}{3}\mathbf{I} - \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}}{r_{ij}^2}\right)\right] \tag{8}$$

where r_{ij} is the distance between beads i and j, and $\mathbf{I}$ is the unit tensor. If beads i and j overlap ($r_{ij} < \sigma_i + \sigma_j$), then

$$\mathbf{D}_{ij} = \frac{k_B T}{6\pi\eta_0\sigma}\left(1 - \frac{9}{32}\frac{r_{ij}}{\sigma}\right)\mathbf{I} + \frac{3}{32}\frac{\mathbf{r}_{ij}\mathbf{r}_{ij}}{r_{ij}\sigma} \tag{9}$$

where $\sigma = \sigma_i = \sigma_j$ if beads are equal size or $\sigma = (\sigma_i + \sigma_j)/2$ otherwise.[59]

Using this representation of the HI effect in the diffusion tensors instead of the original Oseen tensor, the gradient $\partial\mathbf{D}_{ij}/\partial\mathbf{r}_j$ terms in eqs 5 and 7 vanish and the simulation algorithm becomes simpler.

The most time-consuming process in BD with HI is the generation of the random displacement vectors, which require the calculation of a matrix $\mathbf{B}$ that satisfies $\mathbf{D} = \mathbf{B}\cdot\mathbf{B}^T$. For this purpose, McCammon and co-workers[30,46] used Cholesky decomposition, with computing time proportional to N^a with $a = 3$, and Fixman[32] proposed an alternative procedure, based on a Chebyshev polynomial approximation, that has been implemented by some authors,[60−62] with $a \approx 2$. It is clear that for sufficiently large N, the Fixman procedure will be more efficient, although (depending on details of the numerical implementation) for the moderate N employed in many instances, the procedure of McCammon may be faster. BROWFLEX will implement both methods, and a detailed comparison is to be published separately.

As in most dynamic simulation techniques, the time step Δt must be sufficiently small so that the forces do not change much in the step. When hard springs and other strong interactions are present in the model, this requires quite short steps. However, the fluctuations in hydrodynamic interaction are much slower than those interactions, and during such short time steps the change in the diffusion tensor is quite small. Then, in an efficient strategy,[43,63] the $\mathbf{D}$ tensor is not calculated at each time step; instead, it is kept fixed for a block of (say, 5−50) consecutive time steps, during which the same $\mathbf{B}$ is used.

As indicated above, BROWFLEX includes also the simulation of Brownian dynamics in a flowing solvent, which allows the prediction of rheological properties and single-molecule phenomena in flows.[37,39,64] In a homogeneous flow, the velocity of a fluid element can be written as

$$\mathbf{v}_i^0 = \mathbf{G}\cdot\mathbf{r}_i \tag{10}$$

Table 1. Velocity Gradient Tensors for Different Types of Flows

tensor	simple shear	uniaxial elongational	planar elongational
veloc. grad., **G**	$\begin{pmatrix} 0 & \dot{\gamma} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} \dot{\varepsilon} & 0 & 0 \\ 0 & -\dot{\varepsilon}/2 & 0 \\ 0 & 0 & -\dot{\varepsilon}/2 \end{pmatrix}$	$\begin{pmatrix} \dot{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\dot{\varepsilon} \end{pmatrix}$

where $\mathbf{r}_i$ is the position vector of the fluid element, and **G** is the velocity gradient tensor that characterizes the flow. Table 1 gives the expressions[65,66] of that tensor for three common type of flows which, among others, are included in our program.

3. BROWFLEX, the Simulation Program

As commented on above, BROWFLEX is the program devised to perform both equilibrium and nonequilibrium BD simulations of bead-and-spring chains with any topology and with the possibility of selecting among several interaction potentials associated with connectors, angles, torsions, and nonbonded beads.

The information required to control the simulation is organized in several input files having a simple, clear format, so that long input files that are required in some situations can be written by other user-supplied ancillary programs. Thus, as other available software from our lab, BROWFLEX is a data-file driven program, and it is not necessary to have any script for running the simulations. The main input data file just contains the collection of names of both the other input files and the output files, as appreciated in Figure 2A. One of the output files will provide a run-time simulation report and the other one will contain the trajectory, i.e. the Cartesian coordinates defining the macromolecular conformations along the time.

Three compulsory input files are those containing (i) the initial conformation (initcfile.txt) that consists just of a list with the beads Cartesian coordinates; (ii) information on molecular features as number of beads and their radii, connectivity, and parameters of the forces (moleculefile.txt); and (iii) information on simulation features as its duration, the sample size (number of molecules), the time step value, and the type of algorithm used (brownfile.txt). In Figure 2, we show two examples of moleculefile.txt, one for a 12 base pairs double-helical DNA model (Figure 2B), and another one for a 471 base pairs bent DNA model (Figure 2C) (point lines indicate that content is larger but it was omitted to save space). As appreciated, the molecular file is formed by several blocks of information that allow for defining individually the different components of the model. Thus, we have a list with the beads hydrodynamic radii, next a list of bonds where the indices of the two connected beads and the connector force parameters are supplied, next the list of bending interactions with the indices of the three beads involved and the bending force parameters, then a block for torsions that in this case are not present, and finally the list of excluded volume interactions with the indices of the pair of beads not involved in bonds or angles and the excluded volume force parameters.

4. Analysis of Trajectories and Calculation of Properties. ANAFLEX, the Analysis Program

The other program that forms part of the SIMUFLEX suite, named ANAFLEX, was designed to analyze the trajectories generated by BROWFLEX. Separating the generation and the analysis of trajectories has the obvious advantages of speeding up the trajectory generation and allowing for analyzing the trajectory in as many ways as desired.

(A)
```
bentDNA471bp-log.txt          !log file
bentDNA471bp-tra.txt          !trajectory file
moleculefile.txt              !molecular file
initcfile.txt                 !initial file
-                             !no flow file
-                             !no electric field
-                             !no wall file
-                             !no special file
brownfile.txt                 !brownian file
```

(B)
```
20.                     !temperature (Celsius)
0.01                    !solvent viscosity
12000.                  !molecular weight
DNA 20 base pairs       !title
    40                  !number of beads
3.5E-8                  !bead radius
........
   166                  !number of connectors
..............................
 7  10   1   2   1623.9   1.9127E-7 !middle connector
 8  10   1   2   1623.9   1.3584E-7
 9  10   1   2   1623.9   7.0538E-8
10  11   1   2   1623.9   7.0538E-8
10  12   1   2   1623.9   1.3584E-7
10  13   1   2   1623.9   1.9127E-7
10  29   1   2   1623.9   1.9322E-7
10  30   1   2   1623.9   2.0000E-7
10  31   1   2   1623.9   1.9322E-7
..............................
     0              !number of bending angles
     0              !number of torsions
     0              !number of nonbonded interactions
```

(C)
```
20.                       !temperature (Celsius)
0.01                      !solvent viscosity
2.8E+5                    !molecular weight
Bent DNA 471 base pairs   !title
    11                    !number of beads
26.7E-8                   !bead radius
........
    10                    !number of connectors
 1   2   1   2   1.580   160.0E-8     !1st connector
..............................
10  11   1   2   1.580   160.0E-8     !last connector
     9            !number of bending angles
 1   2   3   1   2   0.000   7.39E-14 !1st angle
..............................
 5   6   7   1   2   2.356   7.39E-14 !middle angle
..............................
 9  10  11   1   2   0.000   7.39E-14 !last angle
     0            !number of torsions
     0            !number of nonbonded interactions
```
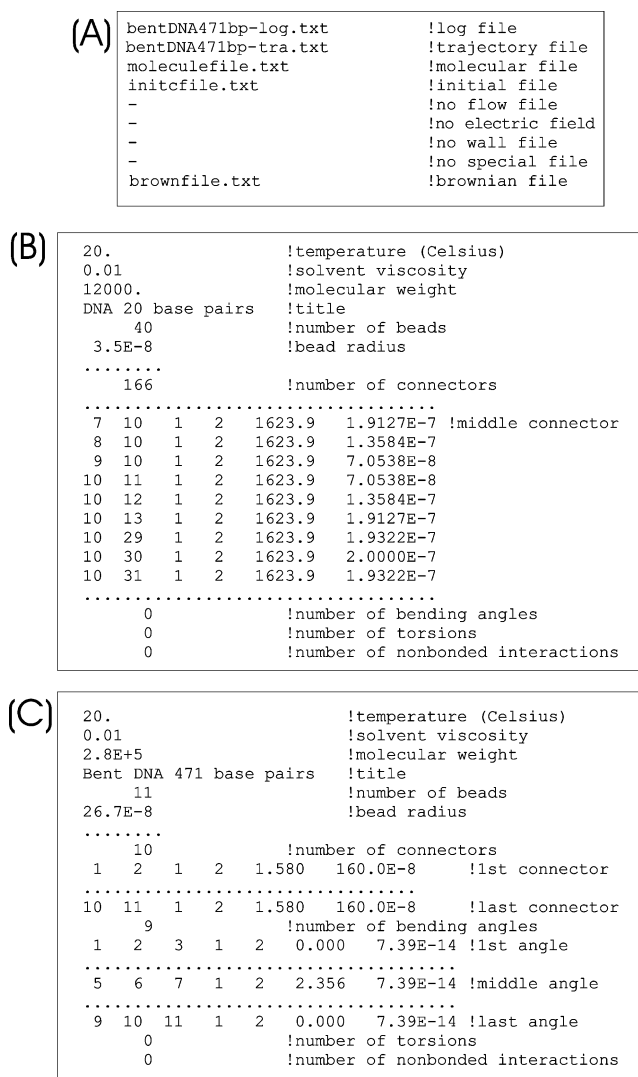
Figure 2. Examples of two of the user-supplied input data files for BROWFLEX: (A) main input file and (B,C) molecular input files for the 12 base pairs double-helical DNA model and for the 471 base pairs bent DNA model (point lines indicate that content is larger but it was omitted to save space).

SIMUFLEX: Algorithms for Flexible Molecules

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2611**

Actually, ANAFLEX analyzes the Brownian trajectories in a number of ways:

• One of the analysis modes consists of the evaluation, as averages over a trajectory of a molecule simulated at equilibrium (steady-state) conditions, of overall properties, either conformational, like the radius of gyration, or hydrodynamic coefficients, in the above-mentioned rigid-body Monte Carlo approach (RBMC),[20−22,67] such as intrinsic viscosity or diffusion coefficient.

• In another mode, the trajectories of a number of molecules can be analyzed obtaining the averages, over the sample, of the properties as a function of time, thus predicting the time evolution of bulk solution properties upon the cessation of external agents (e.g., electrical or flow fields). Of course, it is possible to follow the evolution of each single molecule in order to characterize molecular individualism in single-molecule properties.[39,41,68,69]

• The equilibrium BD trajectories can be also analyzed to study the translational, rotational, and internal dynamics of rigid and flexible particles through the calculation of various time correlation functions, $C(t) = \langle F(t_0, t_0 + t)\rangle_{t_0}$, where $F(t_0, t_0 + t)$ is a quantity that depends on macromolecular conformation at time t_0 and at a later time $t_0 + t$, averaging over all possible choices of the initial time t_0. Some relevant correlation functions computed by ANAFLEX are as follows:

Translational Correlation.

$$C_{trans}(t) = \langle [\mathbf{r}_{cm}(t_0) - \mathbf{r}_{cm}(t_0 + t)]^2\rangle_{t_0} \quad (11)$$

That is the Einstein equation for the center of mass (cm) mean-squared displacement, where the quantity $\mathbf{r}_{cm}$ is the position vector of the center of mass. From a linear fit of $C_{trans}(t)$, the translational diffusion coefficient can be obtained.

Correlation of Any Interelement Vector, Including the End-to-End Vector for Linear Topology.

$$C_{ij}(t) = \langle \mathbf{r}_{ij}(t_0) \cdot \mathbf{r}_{ij}(t_0 + t)\rangle_{t_0} \quad (12)$$

In this case, the correlated quantity is the scalar product of the value of some characteristic vector, $\mathbf{r}_{ij}$, defined between elements i and j of the macromolecular model, at time t_0 times its value at time $t_0 + t$. Similar correlations can be carried out for linear combinations of the $\mathbf{r}_{ij}$'s, such as those involved in the Rouse modes[13,14] of flexible polymer chains. Thus, the longest relaxation time of the chain can be computed from the decay of the correlation function of the first Rouse mode.[28,70] For the specific case of a linear chain, the information on the relaxation time is also contained in the correlation function of the end-to-end vector $\mathbf{r}_{1N}$.

P_2 Function of Some Characteristic Vector. The internal dynamics of a flexible macromolecule can be characterized by the Brownian reorientation of some unitary vector defined inside the molecule, $\mathbf{u}$. In such a case, the correlated quantity will be the angle θ (indeed its cosine) formed by two successive orientations of the vector when a time t has elapsed, i.e. the scalar product of $\mathbf{u}$ at two times separated by t. The correlation functions so defined are the Legendre polynomials. Particularly, a quite common correlation function, involved in the time-dependence of various observable properties (transient electric or flow birefringence and deformation, NMR relaxation, etc.)[30,71] is the second Legendre polynomial, P_2

$$\langle P_2(t)\rangle = \frac{3\langle [\mathbf{u}(t_0) \cdot \mathbf{u}(t_0 + t)]^2\rangle - 1}{2} =$$

$$\frac{3\langle cos^2\theta(t_0, t_0 + t)\rangle_{t_0} - 1}{2} \quad (13)$$

P_2 decay is usually fitted to a multiexponential in order to obtain relaxation times associated with macromolecular internal dynamics. In the case of P_2-related electro-optic properties (birefringence, dichroism...) of rigid macromolecules, theory predicts that a set of up to five reorientational relaxation times can be found.[72,73]

DDLS Correlation Function. Another interesting function based on a second Legendre polynomial is the depolarized dynamic light scattering correlation function, C_{DDLS}[74,75] (also related to the electric birefringence decay that is proportional to the birefringence decay),[76,77] from which information on rigid-body rotation and internal dynamics can be extracted

$$C_{DDLS}(t) = \frac{1}{(N-1)^2} \sum_{j=1}^{N-1} \sum_{j'=1}^{N-1} \langle P_2[\mathbf{u}_j(t_0) \cdot \mathbf{u}_{j'}(t_0 + t)]\rangle_{t_0}$$

$$(14)$$

As observed, C_{DDLS} is related to P_2 and the scalar product $\mathbf{u}_j(t_0) \cdot \mathbf{u}_{j'}(t_0 + t)$ that is the cosine of the angle subtended by the connector vector j at instant t_0 and the connector vector j' at instant $t_0 + t$.

DLS Correlation Function. The (polarized) dynamic light scattering correlation function, C_{DLS}, may allow the simultaneous determination of the translational diffusion coefficient and quantities related to the internal dynamics of flexible particles[28,74,75] and the macromolecular global size

$$C_{DLS}(t, q) = \frac{1}{N^2}\left\langle \sum_{i=1}^{N} \sum_{j=1}^{N} e^{-i\mathbf{q} \cdot [\mathbf{r}_i(t_0) - \mathbf{r}_j(t_0 + t)]}\right\rangle_{t_0} \quad (15)$$

where the modulus of the scattering vector, $|\mathbf{q}| = (4\pi/\lambda)\sin(\theta/2)$, is determined by the wavelength of light, λ, and the scattering angle, θ.

Thus, dynamic coefficients (as translational diffusion coefficient) and time properties (as relaxation times) can be obtained by linear, polynomial, or multiexponential fits of different time correlation functions. Those fits are also carried out by ANAFLEX. In particular, ANAFLEX uses routines adapted from the program DISCRETE[78,79] to make multiexponential fits. It is well-known that fitting multiexponential functions with three or more components is an ill-posed problem. Simpler situations are those of rigid and symmetric particles or weakly bending rods. In any case one can hope to extract the longest or a mean relaxation time.[73]

As BROWFLEX, the ANAFLEX program is driven by simple data files. Figure 3 is an example of the only and simple input file for ANAFLEX. The three first lines are the names of different output files, then it comes a sequence of numeric codes or "flags" to inform ANAFLEX about the

```
bentDNA471bp-res.txt      !results file
bentDNA471bp-log.txt      !log file
bentDNA471bp-sum.txt      !summary file
10                        !sampling frequency
2                         !topology (linear case)
0                         !instant prop.(1-yes)
1                         !analysis mode (steady)
1,6                       !number of properties
18,21,0,0,0               !code for properties
bentDNA471bp-tra.txt      !trajectory file
```

Figure 3. Example of the user-supplied input data file for ANAFLEX.

properties to be calculated and the type of analysis to be performed, and finally the name of the trajectory file is supplied.

5. High-Performance Simulation in Multicore Platforms

In many instances, BD simulation problems are suitable for high-performance computing in multicore servers or clusters, because they may involve somehow independent simulations. The case is trivial when simulation is carried out for each of a sample of many molecules, as in the above-mentioned studies of time-dependency of bulk properties or single molecule behavior. Steady-state averages can be calculated either from a very long trajectory of one molecule or as mean values of the averages of the values computed for a number of molecules, and something similar happens when computing correlation functions $C(t)$. One can either obtain $C(t)$ from a single, very long simulation or determine the function for a number of independent trajectories, averaging $C(t)$ for each time t. The multiple trajectories can be made practically independent, if they are sufficiently long, by changing the sequence of random numbers - in practice, changing the seed of the sequential generator. Truly independent trajectories are those starting from different, initial conformations that would be generated a priori, for instance by Monte Carlo procedures, or even with an inexpensive BD simulation without HI.

In order to take advantage of multicore computers and clusters (even a computer with two *Quad* processor has eight cores), we have set up a scheme to run such multimolecule simulations, based on two ancillary tools. First, Multi-BROWFLEX "clones" what would be the files for a single simulation, producing multiple copies changing either the seed of the random numbers or the file containing the initial coordinates. It also generates a batch file for all the execution that is submitted to a load-balancing manager such as Sun Grid Engine. The outcome consists of multiple results files, one for each trajectory. Then, there is another tool, Multi-ANAFLEX, which is in charge of collecting and reading all those files, producing the final results as the proper averages over those of each molecule.

6. Examples: Multiscale Simulations of DNA

In order to demonstrate the usefulness and versatility of the mechanical model, BD algorithms and other methodologies implemented in the BROWFLEX suite, we have chosen a well-known and most relevant macromolecule, DNA, in sizes ranging from ~10 through ~10^5 base pairs. With a convenient, more or less coarse-grained bead-and-connector model, simulations of the dynamics in such multiple scales are possible employing the same methodology. Next, we show how BROWFLEX works by applying it to study the dynamics of several DNA models, always in experimentally observable situations.

6.1. A Double-Helical Model. There are macromolecular solution properties related to the local dynamics and structure of the molecule. In case of B-DNA this implies the convenience of modeling the double-helix properly. It is clear that atomic level simulations are quite expensive in CPU time. Then a suitable mesoscale model is built by considering the nucleotides in each strand as repetitive units. That model, proposed years ago by Horta and García de la Torre,[80,81] contains the characteristic double-helix parameters: number of base pairs, pitch, phase angle, helix diameter, etc. This is a highly valuable model to represent short fragments of DNA. Our group has already employed that model, where for simplicity beads lie only on the outside, in order to study hydrodynamic properties of double-helical DNA[63] (later, in the spirit of this kind of mesoscale simulations, similar models have been employed by other workers).[82,83] If larger DNA fragments were to be simulated, possible interpenetration of double helices can be avoided by using the hard-sphere excluded volume included in our program or alternatively some other DNA model.[84,85]

In this model, nucleotides are modeled by beads all of them with the same hydrodynamic radius, σ. Then, the number of beads will be $N = 2N_{bp}$, where N_{bp} is the number of base pairs (bp). In addition, some degree of flexibility is incorporated by using hard Hookean, elastic connectors (Fraenkel springs, defined in paragraph after eq 2). For the sake of minimizing the amount of interactions but keeping the double-helical shape and the stiffness at short scale, we found it adequate to connect each bead i to

1. its first neighbors along its strand (beads $i \pm 1$), which keeps connectivity and bond equilibrium length.

2. its second neighbors along its strand (beads $i \pm 2$), which accounts for bending interactions.

3. its third neighbors along its strand (beads $i \pm 3$), which accounts for torsional interactions.

4. its counterpart in the other strand (bead $i + N_{bp}$), which accounts for interactions between nucleotides forming the base pair.

5. the first neighbors of its counterpart in the other strand (beads $i + (N_{bp} \pm 1)$), which is necessary in order to keep the strands together.

Figure 4 shows this model (in the straight, equilibrium conformation) displaying all the connectors involving one of the innermost beads (indeed, these are the connectors whose data have been kept in Figure 2B).

As described in ref 86, we performed BD simulations by using program BROWFLEX for double-helical DNA models representing oligonucleotides with a different number of base pairs N_{bp}, all of them with the following features: helix radius, $r = 10$ Å, pitch = 3.4 Å, phase angle either $\phi = 180°$ (a symmetrical helix, Figure 4) or $\phi = 120°$ (a nonsymmetrical helix, more akin to the Watson–Crick structure), $H =$

SIMUFLEX: Algorithms for Flexible Molecules

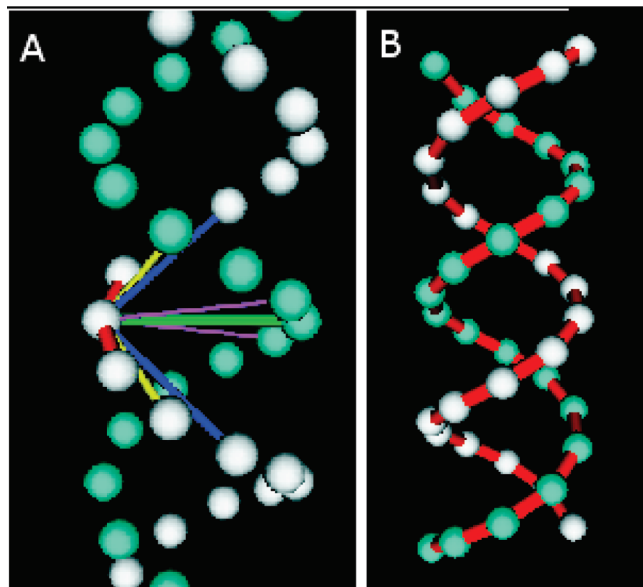*J. Chem. Theory Comput., Vol. 5, No. 10, 2009* **2613**

Figure 4. Double-helical model for DNA (A) showing the connectors supported by one bead and (B) showing only the connectors between neighbors in each strand.

Figure 5. Sketch of a nearly touching beads model where each bead embraces a piece of the DNA double helix.

Table 2. Diffusion Coefficients and Orientational Relaxation Times of DNA Oligonucleotides Obtained via Brownian Dynamics Simulations[a]

N_{bp}	simul. $\phi = 180^0$	simul. $\phi = 120^0$	exptl
	$D_t \times 10^7 \text{(cm}^2 \text{ s}^{-1})$		
8	16.3 ± 1.8	16.7 ± 1.6	15.3
12	13.2 ± 0.4	13.4 ± 1.2	13.4
20	10.6 ± 0.3	10.8 ± 0.5	10.9
	τ_{ee} (ns)		
8	3.3 ± 0.2	3.9 ± 0.5	3.2
12	6.1 ± 0.7	7.8 ± 0.8	6.4
20	16.7 ± 1.7	19.0 ± 1.3	16.2

[a] Comparison to experimental values obtained by Eimer and Pecora.[87]

$200k_BT/b_1^2$, where b_1 is the equilibrium length of the connector binding bead i to its first neighbor in its own strand, $T = 293$ K, $\eta_s = 0.01$ poise, hydrodynamic bead radius $\sigma = 3.5$ Å, and equilibrium spring length $b_1 = 7.0$ Å. The trajectories generated by the Brownian dynamics simulation were analyzed with program ANAFLEX in order to compute the translational diffusion coefficient, D_t, and the rotational relaxation time of the end-to-end vector, τ_{ee}. Within the statistical uncertainty of the simulations, the results for the two choices of ϕ are identical. Table 2 shows a comparison between the values of those dynamic properties obtained by simulation and the experimental values obtained by Eimer and Pecora.[87] As appreciated, the agreement is quite good. It is noteworthy that D_t and τ_{ee} are indeed quite close to the predictions of a rigid-body hydrodynamic calculation, using the HYDRO++ program[8] for the straight equilibrium conformation, demonstrating, thanks to BD simulations, the validity of the RBMC treatment for overall properties of quite stiff molecules. However, the relaxation time for the $\langle P_2(t) \rangle$ function for a vector perpendicular to the helical axis differs remarkably from the rigid-body prediction. While bending is scarcely noticeable in such short oligonucleotides, torsion of the helix, which influences the diffusivity of such
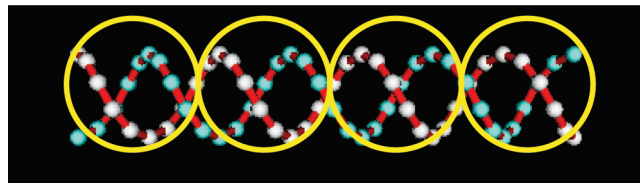
transversal vector, is much more influential (for more details, see refs 63 and 86).

Mesoscale models, with one, two, or a few elements per amino acid residue[15,88] are now being considered as an alternative to atomistic simulation to predict dynamics of natively unfolded (intrinsically disordered) proteins or that of the folding process. Such models, and the important, recently demonstrated,[89] HI effects in protein dynamics, can be very adequately covered with our methodology.

6.2. A String of Touching Beads. A higher level in the coarse-graining procedure results from considering the macromolecular chain from a global perspective for which the precise local structure is not relevant and assuming that its flexibility is more or less uniformly distributed along its backbone. These assumptions give rise to the well-known wormlike chain model (or Kratky−Porod chain). In such a model, the double-helical structure of B-DNA is not explicitly considered. The wormlike chain can be represented with the generalized bead-and-spring model by setting constant the connector lengths, as in the case of a freely jointed chain, and allowing for the bond angles to fluctuate around their equilibrium values (0° for an actual wormlike chain) with an amplitude that depends on the flexibility of the chain. Thus, the model can be used to represent from rigid to flexible structures just by playing with the bond angle parameters. In the same way, the model can be used to represent structures with local bents just by setting the bond angle located at the bent position to its characteristic value.

A limiting case of the wormlike chain model would be a string of nearly touching beads each embracing a piece of the double helix (see Figure 5). This macromolecular representation, which could be termed as a "fine-grained" model, was initially suggested by Schellman,[90] implemented in Monte Carlo simulations by Hagerman and Zimm[22] and in Brownian dynamics simulation by Allison and McCammon,[30,91] and is useful in representing short fragments of DNA of a few hundreds of base pairs. The bead diameter, which is the same as the connector length, is set to $b = 24.5$ Å. In that way, the diffusion coefficients of a straight string of beads are practically identical to those of a cylinder with a diameter of 20 Å, which is the hydrated diameter of DNA deduced from cylindrical models.[92] Then, the contour length of a model chain with N beads would be $L = Nb = 24.5N$ Å. On the other hand, the length of a double-helical B-DNA related to the number of base pairs is $L = 3.4N_{bp}$ Å so that the relationship between N and N_{bp} is $N = 0.14N_{bp}$. In the Schellman-Hagerman-Zimm model, the flexibility of the chain is represented by a bending potential quadratic in the bending angle, α, subtended by two successive links between neighbor beads. The bending force constant is related to the persistence length, P, by $Q = k_BTP/b$. Therefore,
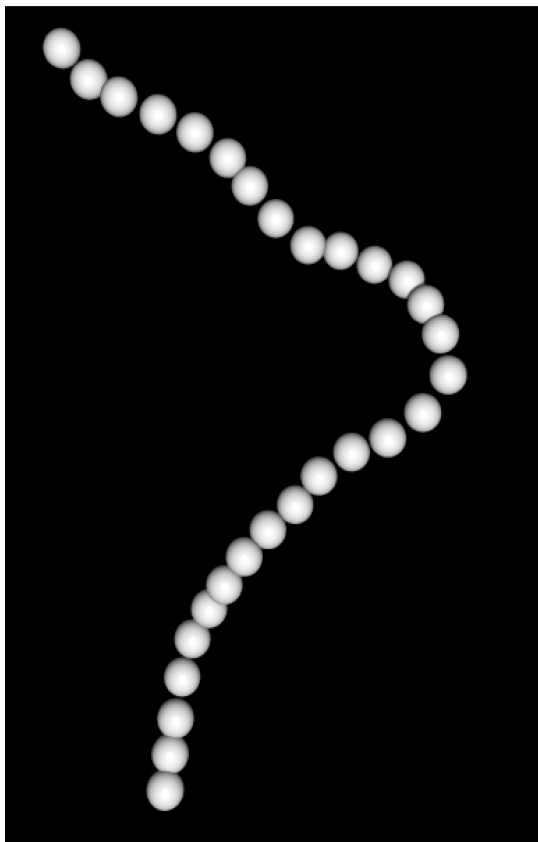
Figure 6. Nearly touching beads model for a DNA of 207 base pairs.

for a typical $P = 500$ Å and a $T = 293$ K, $Q = 8.24 \times 10^{-13}$ erg. The connector length is kept more or less constant ($\sim$10% in length fluctuation) by using stiff Fraenkel springs with a spring constant $H = 100k_BT/b^2 = 67.4$ erg/cm^2.

According to the above specifications and values of the model parameters, a DNA of 207 bp was modeled by 29 touching beads as illustrated in Figure 6. We simulated two variants of such a DNA molecule: a) an unbent DNA, with the equilibrium value of all the bond angles equal to zero, and b) a bent DNA, with the equilibrium value of the bond angles zero except for the central one that was set to 40°.

This comparison is intended to analyze the effect of such sharp bents, induced by some special base sequences, in the diffusivity of short DNA molecules.[73,93] For both model chain, BD simulations with HI were carried out by using BROWFLEX. From the generated trajectories, the translational diffusion coefficient and the orientational longest relaxation time of the end-to-end vector coming from both the correlation function based on depolarized dynamic light scattering (applicable to electric birefringence decay) and the correlation function based on the second Legendre polynomial P_2 were computed by using ANAFLEX. As noted above, Hagerman and Zimm[22] anticipated, other works[67,94] confirmed that the RBMC treatment (implemented in MONTEHYDRO)[19] works well for quite stiff semiflexible macromolecules in the prediction of not only translational diffusion but also overall rotational diffusion. Table 3 demonstrates that the RBMC results are in very good agreement with those of the BD simulations.

Table 3. Diffusion Coefficients and Orientational Relaxation Times of Bent and Unbent DNA Obtained via Brownian Dynamics and Monte Carlo Simulations

	bent	unbent
$D_t \times 10^7$ (cm^2 s^{-1}) (RBMC)	2.35	2.33
$D_t \times 10^7$ (cm^2 s^{-1}) (BD)	2.34	2.32
τ_{ee} (μs) (BD-P$_2$)	2.8	3.2
τ_{ee} (μs) (BD-DDLS)	2.5	3.1

6.3. Coarse-Grained Model of the Wormlike Chain. A cruder coarse-grained representation of the wormlike chain mentioned above consists of a string of nontouching beads connected by stiff springs, including a bending potential between successive connectors that determines the persistence length of the chain (see Figure 7). BD simulations on this kind of model were presented years ago by Allison and co-workers.[31,76,91,95]

The only free parameter of this model is the number of beads, N. As long as N is large enough the results are independent of its exact value. Thus, we can assign values to the parameters of our bead-and-spring chain by previously choosing a value for N (it is remarkable that a DNA molecule with 2311 base pairs can be modeled as a chain of only 10 beads).[95] Then, the connector length is fixed by the relationship $b = L/(N - 1)$. The constant of the stiff springs is set to $H = 100k_BT/b^2$, and the equilibrium bending angle is set to $\alpha_0 = 0$. Finally, the constant of the bending potential, Q, is chosen to adjust the persistence length, P, or the radius of gyration, R_g, of the real macromolecular chain, and the bead radius, σ, is chosen to adjust its translational diffusion coefficient, D_t. Using the just described parametrization procedure, we modeled an unbent DNA of 471 bp with $N = 21$ beads and a bent DNA of 471 bp with both $N = 11$ and $N = 21$ beads. The central bent was set to 45° in order to reproduce the kind DNA worked out by Stellwagen and co-workers.[93] Figure 8 shows the evolution of the depolarized dynamic light scattering correlation function for the three simulated DNA model chains. Those correlation functions were obtained after analyzing with ANAFLEX the corresponding Brownian trajectories with a duration $t = 2000\ \mu$s (much longer than the longest relaxation time) generated with BROWFLEX. As observed, the value of N does not influence the results. On the other hand, the difference in dynamics of bent and unbent DNA is easily characterized. That figure is in agreement with Figure 2(b) in ref 93.

6.4. Single-Molecule Stretching of Long DNA in a Flow Field. It is well-known that flexible polymer chains subjected to extensional flows with a rate of strain greater than a certain critical value experience the so-called "coil-
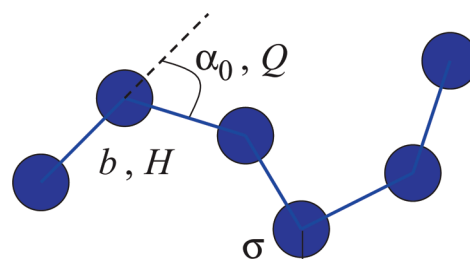

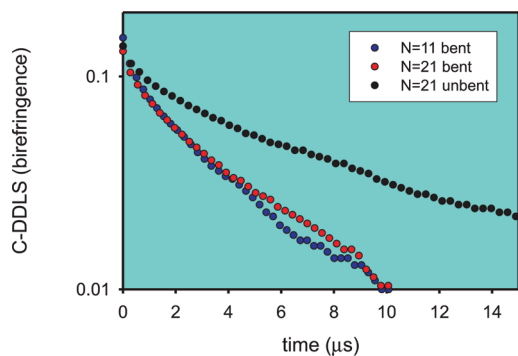***Figure 7.*** Bead-and-connector model for a wormlike chain.

Figure 8. Evolution of the DDLS correlation function for the three DNA molecules modeled as wormlike chains.

Figure 9. Time evolution of the extension of individual DNA molecules subjected to elongational flow. Comparison of experimental (left graph adapted from ref 41) and simulation (right graph) results.

Figure 10. Time evolution of the distribution of the extension of a sample of DNA molecules. Comparison of experimental (left graph adapted from ref 41) and simulation (right graph) results.

stretch" transition.[96] This phenomenon consists of the abrupt unraveling of the random coil to a stretched conformation. In an already classical series of single-molecule experiments carried out with DNA, Chu and co-workers showed that the coil−stretch transition occurs in a particular manner for each chain in a given sample.[41,97,98] This is called molecular individualism.[68] Brownian dynamics simulation has been revealed to be an adequate technique to reproduce such a behavior.[17,39,40]

We show that BROWFLEX is able to reproduce the experimental results of a sample of DNA molecules as those employed in one of the pioneering experiments by Chu et al.[41] In that experiment a circular λ-DNA with 48.5×10^3 base pairs was converted into a linear chain by thermal treatment. Then, each macromolecular chain was stained fluorescently, subjected to elongational flow, and its stretching behavior was monitored by means of fluorescence microscopy. According to the authors of that work the stained DNA employed had a contour length $L = 22 \, \mu$m and a radius of gyration $\langle s^2 \rangle^{1/2} = 0.73 \, \mu$m (as estimated from the measured translational diffusion coefficient, $D_t = 0.47 \, \mu$m^2/s). Thus, assuming a characteristic persistence length for DNA of $P = 0.05 \, \mu$m, the ratio $L/P \simeq 440$ ensures that DNA is a flexible chain with random coil equilibrium conformation. Finally authors inform that the solvent viscosity is $\eta_s = 41$ cP and the working temperature is $T = 22.7$ °C. Since we are interested in reproducing the global conformation and dynamics of a quite large and flexible DNA chain, we can use a rough coarse-grained model. The model consists of a linear chain of $N = 20$ beads connected by $N - 1 = 19$ FENE springs that are able to capture both the Gaussian statistic that appears at low strain rate and the finite extensibility that plays a role at high strain rate. Thus, each bead represents a large DNA segment. Taking into account that $L = l_{max}(N - 1)$, we get a maximum spring length $l_{max} = 1.16 \, \mu$m. Then, by using the experimental radius of gyration, we get a value for the equilibrium spring length that, after some fitting refinements, turns out to be $b = 0.448 \, \mu$m.

We performed BD simulations of that FENE chain under elongational flow without EV interactions, which implies theta conditions, and including fluctuating HI, with a value of the hydrodynamic parameter $h^* = 0.25$, which corresponds to a hydrodynamic bead radius $\sigma = 0.257b$.
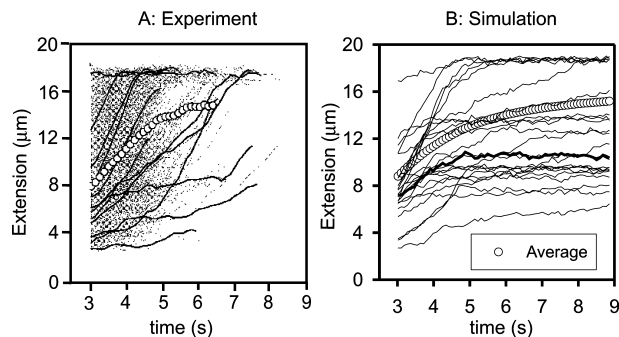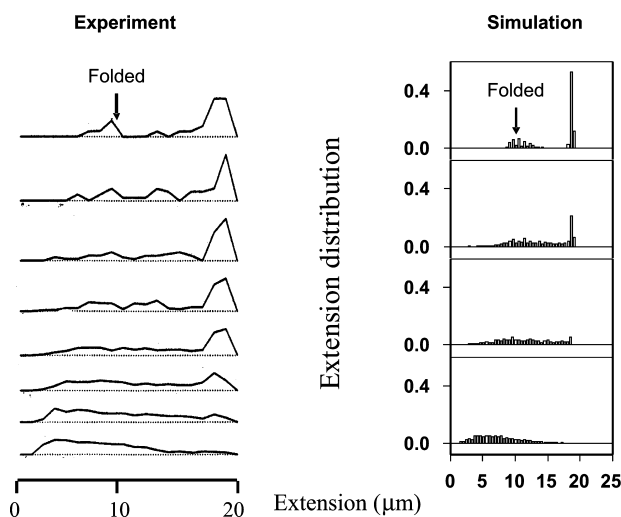
One of the optional input files, intended to describe steady or time-dependent flows (see BROWFLEX user guide), is used to specify a steady elongational rate $\dot{\varepsilon} = 0.86$ s^{-1}. That is the maximum elongational rate employed in ref 41, a value for which the molecular individualism is more clearly appreciated. The property monitored in the experiments by Chu and co-workers was the chain extension along the flow direction; therefore, we studied also the time evolution of that chain property. In order to get a good ensemble, we simulated 1500 molecules. In this study, ANAFLEX works in the multimolecule mode, supplying the time evolution of the sample-average and single-molecule properties (the extension as measured by Chu and co-workers is among the numerous properties that the program can analyze).

Figure 9 is a comparison of the evolution of the molecular extension computed from our simulations to that obtained experimentally.[41] The molecular individualism is appreciated as the particular paths followed by the time evolution of the chain extension: different chains experience coil−stretch transition at different times and reach different amount of extension. As observed, simulation results are in agreement with experiments as well as with other computer simulations.[99] Furthermore, in Figure 10 we compare the time

evolution of the histogram representing the probability distribution of the molecular extension obtained from simulations with that reported by Perkins et al. (Figure 1(B) in ref 41). Both histograms series are in excellent agreement. Initially, when the Hencky strain is small $\dot{\varepsilon}t = 2.5$, most of the chains are close to their coil conformation although some of them can be stretched, and therefore both histograms are slightly broad and present a maximum at low extension (~ 5 μm). As time goes by and strain increases, histograms start to get broader and eventually a peak at a extension corresponding to the fully stretched conformation starts to develop owing to the increase number of chains that become stretched. That peak occurs at the same strain value in both experiment and simulation. Interestingly, at a higher strain a second peak at half the maximum extension arises. That second peak corresponds to folded "hairpin" conformations, which have an extension approximately half of the contour length. Again, the simulation was able to reproduce the experimental evidence.

7. Concluding Remarks

We provide a useful tool for the Brownian dynamics simulation and analysis of flexible and semiflexible bead-and-spring macromolecular chain models, SIMUFLEX, which consists of two programs, BROWFLEX and ANAFLEX. The programs are easy to use and were designed to be quite general. Thus, BROWFLEX can handle macromolecular models with any topology and include a number of common interaction potentials that can be easily extended in future versions. A key feature of this BD simulation tool is the inclusion of fluctuating HI, that allows to perform more realistic simulations. On the other hand, ANAFLEX is quite simple to employ and contemplates the analysis of a number of solution properties (in both steady-state and time-dependent conditions) and time correlation functions, which can also be extended in future versions. The examples employed in this paper have tried to show how the programs work and have in common to correspond to simulations of DNA dynamics in different scales. Thus, it was shown how SIMUFLEX is able to produce results comparable to experiments for different DNA problems that require to model the DNA chain with a different coarse-grain level.

8. Computer Methods

The SIMUFLEX suite that includes the BROWFLEX and ANAFLEX programs, as well as MONTEHYDRO and other related public-domain software, can be freely downloaded from our Web site at http://leonardo.inf.um.es/macromol.

References

(1) Teraoka, I. *Polymer Solutions*; John Wiley and Sons: New York, 2002.

(2) Serdyuk, I. N.; Zaccai, N. R.; Zaccai, J. *Methods in Molecular Biophysics. Structure, Dynamics, Function*; Cambridge University Press: New York, 2007.

(3) Schlick, T. *Molecular Modeling and Simulation*; Springer: New York, 2002.

(4) García de la Torre, J.; Bloomfield, V. A. *Q. Rev. Biophys.* **1981**, *14*, 81–39.

(5) García de la Torre, J.; Navarro, S.; López Martínez, M. C.; Díaz, F. G.; López Cascales, J. J. *Biophys. J.* **1994**, *67*, 530–531.

(6) Carrasco, B.; García de la Torre, J. *Biophys. J.* **1999**, *76*, 3044–3057.

(7) García de la Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719–730.

(8) García de la Torre, J.; Del Río Echenique, G.; Ortega, A. *J. Phys. Chem. B* **2007**, *111*, 955–961.

(9) Allison, S. A. *Biophys. Chem.* **2001**, *93*, 197–213.

(10) Hahn, D. K.; Aragon, S. R. *J. Chem. Theory Comput.* **2006**, *2*, 1416–1428.

(11) Mansfield, M. L.; Douglas, J. F.; Kang, E. H. *Macromolecules* **2007**, *40*, 2575–2589.

(12) Allison, S. A.; Pei, H.; Haynes, M.; Xin, Y.; Law, L.; Labrum, J.; Augustin, D. *J. Phys. Chem. B* **2009**, *112*, 5858–5866.

(13) Rouse, P. E. *J. Chem. Phys.* **1953**, *21*, 1272–1280.

(14) Doi, M.; Edwards, S. F. *The Theory of Polymer Dynamics*; Oxford University Press: Oxford, 1986.

(15) Ayton, G. S.; Noid, W. G.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(16) Prabhakar, R.; Prakash, J. R.; Sridhar, T. *J. Rheol.* **2004**, *48*, 1251–1278.

(17) Sunthar, P.; Prakash, J. R. *Macromolecules* **2005**, *38*, 617–640.

(18) García de la Torre, J.; Pérez Sánchez, H. E.; Ortega, A.; Hernández Cifre, J. G.; Fernandes, M. X.; Díaz Baños, F. G.; López Martínez, M. C. *Eur. Biophys. J* **2003**, *32*, 477–486.

(19) García de la Torre, J.; Ortega, A.; Pérez Sánchez, H. E.; Hernández Cifre, J. G. *Biophys. Chem* **2005**, *116*, 121–128.

(20) Zimm, B. H. *Macromolecules* **1980**, *13*, 592–602.

(21) García de la Torre, J.; Jiménez, A.; Freire, J. J. *Macromolecules* **1982**, *15*, 148–154.

(22) Hagerman, P.; Zimm, B. H. *Biopolymers* **1981**, *20*, 1481–1502.

(23) Harvey, S. H.; García de la Torre, J. *Macromolecules* **1980**, *13*, 960–964.

(24) Allen, M. P.; Tildsley, D. J. *Computer Simulation of Liquids*, 1st ed.; Clarendon: Oxford, 1987.

(25) Díaz, F. G.; Iniesta, A.; García de la Torre, J. *J. Chem. Phys.* **1987**, *87*, 6021–6027.

(26) Díaz, F. G.; García de la Torre, J. *J. Chem. Phys.* **1988**, *88*, 7698–7705.

(27) Díaz, F. G.; García de la Torre, J.; Freire, J. J. *Polymer* **1989**, *30*, 259–264.

SIMUFLEX: Algorithms for Flexible Molecules

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2617**

(28) Rey, A.; Freire, J. J.; García de la Torre, J. *J. Chem. Phys.* **1989**, *90*, 2035–2041.

(29) Iniesta, A.; García de la Torre, J. *J. Chem. Phys.* **1990**, *92*, 2015–2019.

(30) Allison, S. A.; McCammon, J. A. *Biopolymers* **1984**, *23*, 167–187.

(31) Allison, S. A.; McCammon, J. A. *Biopolymers* **1984**, *23*, 363–375.

(32) Fixman, M. *Macromolecules* **1986**, *19*, 1195–1204.

(33) Fixman, M. *Macromolecules* **1986**, *19*, 1204–1207.

(34) Zylka, W.; Öttinger, H. C. *J. Chem. Phys.* **1989**, *90*, 474–480.

(35) Yamakawa, H. *Modern Theory of Polymer Solutions*; Harper & Row: New York, 1971.

(36) García de la Torre, J.; López Martínez, M. C.; Tirado, M. M.; Freire, J. J. *Macromolecules* **1983**, *16*, 1121–1127.

(37) Hernández Cifre, J. G.; García de la Torre, J. *J. Rheol.* **1999**, *43*, 339–358.

(38) Rossky, P. J.; Doll, J. D.; Friedman, H. L. *J. Chem. Phys.* **1978**, *69*, 4628–4633.

(39) Larson, R. G.; Hu, H.; Smith, D. E.; Chu, S. *J. Rheol.* **1999**, *43*, 267–304.

(40) Hur, J. S.; Shaqfeh, E. S. G.; Larson, R. G. *J. Rheol.* **2000**, *44*, 713–742.

(41) Perkins, T. T.; Smith, D. E.; Chu, S. *Science* **1997**, *276*, 2016–2021.

(42) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(43) Klenin, K.; Merlitz, H.; Langowski, J. *Biophys. J.* **1998**, *74*, 780–788.

(44) Del Río Echenique, G.; Rodríguez Schmidt, R.; Freire, J. J.; Hernández Cifre, J. G.; García de la Torre, J. *J. Am. Chem. Soc.* **2009**, *131*, 8548–8556.

(45) Fraenkel, G. K. *J. Chem. Phys.* **1952**, *20*, 642–647.

(46) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352–1360.

(47) Warner, H. R. *Ind. Eng. Chem. Fundam.* **1972**, *11*, 379–387.

(48) Pamies, R.; Hernández Cifre, J. G.; García de la Torre, J. *J. Polym. Sci. Pol. Phys.* **2007**, *45*, 1–9.

(49) Carrasco, B.; Pérez Belmonte, A.; López Martínez, M. C.; García de la Torre, J. *J. Chem. Phys.* **1996**, *100*, 9900–9905.

(50) Navarro, S.; Carrasco, B.; López Martínez, M. C.; García de la Torre, J. *J. Polym. Sci. Pol. Phys.* **1997**, *35*, 689–697.

(51) Fernandes, M. X.; Castanho, M. A. R. B.; García de la Torre, J. *BBA-Biomembranes* **2002**, *1565*, 29–35.

(52) Chirico, G.; Langowski, J. *Macromolecules* **1992**, *25*, 769–775.

(53) Öttinger, H. C. *Stochastic Processes in Polymer Fluids*; Springer: Berlin, 1996.

(54) Jian, H. M.; Vologodskii, A. V.; Schlick, T. *J. Comput. Phys.* **1997**, *136*, 168–179.

(55) Benham, C. J.; Mielke, S. P. *Annu. Rev. Biomed. Eng.* **2005**, *7*, 21–53.

(56) Rotne, J.; Prager, S. *J. Chem. Phys.* **1969**, *50*, 4831–4837.

(57) Yamakawa, H. *J. Chem. Phys.* **1970**, *53*, 436–443.

(58) García de la Torre, J.; Bloomfield, V. A. *Biopolymers* **1977**, *16*, 1747–1763.

(59) Carrasco, B.; García de la Torre, J.; Zipper, P. *Eur. Biophys. J.* **1999**, *28*, 510–515.

(60) Kröger, M.; Alba Pérez, A.; Laso, M.; Öttinger, H. C. *J. Chem. Phys.* **2000**, *113*, 4767–4773.

(61) Jendrejack, R. M.; Graham, M. D.; de Pablo, J. J. *J. Chem. Phys.* **2000**, *113*, 2894–2900.

(62) Schlick, T.; Beard, D. A.; Huang, J.; Strahs, D. A.; Qian, X. *Comput. Sci. Eng.* **2000**, *2*, 38–51.

(63) Huertas, M. L.; Navarro, S.; López Martínez, M. C.; García de la Torre, J. *Biophys. J.* **1997**, *73*, 3142–3153.

(64) Pamies, R.; López Martínez, M. C.; Hernández Cifre, J. G.; García de la Torre, J. *Macromolecules* **2005**, *38*, 1371–1377.

(65) Macosko, C. W. *Rheology. Principles, measurements and applications*; VCH Publishers: New York, 1994.

(66) Larson, R. G. *The Structure and Rheology of Complex Fluids*; Oxford University Press: New York, 1999.

(67) Iniesta, A.; Díaz, F. G.; García de la Torre, J. *Biophys. J.* **1988**, *54*, 269–275.

(68) De Gennes, P. G. *Science* **1997**, *276*, 1999.

(69) Hernández Cifre, J. G.; García de la Torre, J. *J. Non-Cryst. Solids* **1998**, *235−237*, 717–722.

(70) Rey, A.; Freire, J. J.; García de la Torre, J. *Macromolecules* **1990**, *23*, 3948–3953.

(71) Allison, S. A. *J. Chem. Phys.* **1989**, *90*, 3843–3849.

(72) Favro, L. D. *Phys. Rev.* **1960**, *119*, 53–62.

(73) García de la Torre, J. *Colloids Surf. B* **2007**, *56*, 4–15.

(74) Berne, B.; Pecora, R. *Dynamic Light Scattering*; John Wiley and Sons: New York, 1976.

(75) Chu, S. In *Soft Matter Characterization*; Borsali, R., Pecora, R., Eds.; Springer: New York, 2008; Chapter 7.

(76) Lewis, R. J.; Allison, S. A.; Eden, D.; Pecora, R. *J. Chem. Phys.* **1988**, *89*, 2490–2503.

(77) Allison, S. A.; Nambi, P. *Macromolecules* **1992**, *25*, 759–769.

(78) Provencher, S. *J. Chem. Phys.* **1976**, *64*, 2772–2777.

(79) Provencher, S. *Biophys. J.* **1976**, *16*, 151–170.

(80) García de la Torre, J.; Horta, A. *J. Phys. Chem.* **1976**, *80*, 2028–2035.

(81) García de la Torre, J.; Horta, A. *Polym. J.* **1977**, *9*, 33–39.

(82) Drukker, K.; Wu, G.; Schatz, G. C. *J. Chem. Phys.* **2001**, *114*, 579–588.

(83) Tepper, H.; Voth, G. A. *J. Chem. Phys.* **2005**, *122*, 124906.

(84) Malhorta, A.; Tan, R. K.-Z.; Harvey, S. C. *Biophys. J.* **1994**, *66*, 1777–1795.

(85) Devkota, B.; Petrov, A. S.; Lemieux, S.; Boz, M. B.; Tang, L.; Schneemann, A.; Johnson, J. E.; Harvey, S. C. *Biopolymers* **2009**, *91*, 530–538.

(86) García de la Torre, J.; Navarro, S.; López Martínez, M. C. *Biophys. J.* **1994**, *66*, 1573–1579.

(87) Eimer, W.; Pecora, R. *J. Chem. Phys.* **1991**, *94*, 2324–2329.

(88) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(89) Frembgen-Kesner, T.; Elcock, A. H. *J. Chem. Theory Comput.* **2009**, *5*, 242–256.

(90) Schellman, J. A. *Biopolymers* **1974**, *13*, 217–226.

(91) Allison, S. A. *Macromolecules* **1986**, *19*, 118–124.

(92) Tirado, M. M.; López Martínez, M. C.; García de la Torre, J. *J. Chem. Phys.* **1984**, *81*, 2047–2052.

(93) Lu, Y.; Weers, B. D.; Stellwagen, N. *Biophys. J.* **2003**, *85*, 409–415.

(94) García de la Torre, J. *Eur. Biophys. J.* **1994**, *23*, 307–322.

(95) Allison, S. A.; Sorlie, S. S.; Pecora, R. *Macromolecules* **1990**, *23*, 1110–1118.

(96) De Gennes, P. G. *J. Chem. Phys.* **1974**, *60*, 5030–5042.

(97) Smith, D. E.; Chu, S. *Science* **1998**, *281*, 1335–1340.

(98) Smith, D. E.; Babcock, H. P.; Chu, S. *Science* **1999**, *283*, 1724–1727.

(99) Schroeder, C. M.; Shaqfeh, E. S. G.; Chu, S. *Macromolecules* **2004**, *37*, 9242–9256.

JCTC Journal of Chemical Theory and Computation

Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics

Ivan S. Ufimtsev and Todd J. Martinez*

Department of Chemistry, Stanford University, Stanford, California 94305

Abstract: We demonstrate that a video gaming machine containing two consumer graphical cards can outpace a state-of-the-art quad-core processor workstation by a factor of more than $180\times$ in Hartree−Fock energy + gradient calculations. Such performance makes it possible to run large scale Hartree−Fock and Density Functional Theory calculations, which typically require hundreds of traditional processor cores, on a single workstation. Benchmark Born−Oppenheimer molecular dynamics simulations are performed on two molecular systems using the 3-21G basis set - a hydronium ion solvated by 30 waters (94 atoms, 405 basis functions) and an aspartic acid molecule solvated by 147 waters (457 atoms, 2014 basis functions). Our GPU implementation can perform 27 ps/day and 0.7 ps/day of *ab initio* molecular dynamics simulation on a single desktop computer for these systems.

Introduction

The idea of using graphical hardware for general purpose computing goes back over a decade.[1−3] Nevertheless, early attempts to use GPUs for scientific calculations were largely stymied by lack of programmability and low precision of the hardware. The recent introduction of the Compute Unified Device Architecture (CUDA) programming interface[4] and hardware capable of performing double precision arithmetic operations by Nvidia significantly simplified GPU programming and has triggered an increasing number of publications in different fields, such as classical molecular dynamics[5−9] and quantum chemistry.[10−15] Our initial implementation of two-electron repulsion integral evaluation algorithms[13] and the entire direct self-consistent field procedure on the GPU[14,15] demonstrated the large potential of graphical hardware for quantum chemistry calculations. In this article, we continue exploring the use of GPUs for quantum chemistry, including the calculation of analytical gradients for self-consistent field wave functions and the implementation of *ab initio* Born−Oppenheimer (BO) molecular dynamics. We compare the GPU performance results to the GAMESS[22] quantum chemistry package running on an Intel Core2 quad-core 2.66 GHz CPU workstation, a state-of-the-

art desktop computing system. Our "machine to machine" rather than "GPU to CPU core" comparison provides a realistic estimate of the actual speedup that can be obtained in real calculations. Comparison of the total time required to calculate energies and gradients on these two platforms shows that the GPU workstation is 2 orders of magnitude faster than the CPU workstation. This allows us to carry out *ab initio* molecular dynamics of large systems at more than a thousand MD steps per day on a desktop computer.

Analytical Energy Gradient Implementation. The general formula for nuclear gradients follows directly from the expression for the Hartree−Fock energy including a term accounting for the basis set dependence on molecular geometry[16]

$$\nabla_A E_{HF} = \sum_{\mu\nu} D_{\mu\nu}(\nabla_A H_{\mu\nu}) - \sum_{\mu\nu} W_{\mu\nu}(\nabla_A S_{\mu\nu}) + \sum_{\mu\nu\lambda\sigma}\left(D_{\mu\nu}D_{\lambda\sigma} - \frac{1}{2}D_{\mu\lambda}D_{\nu\sigma}\right)[\nabla_A(\mu\nu)|\lambda\sigma] \quad (1)$$

where A labels an atomic center, $W_{\mu\nu}$ is the energy weighted density matrix, $S_{\mu\nu}$ is the overlap matrix, $D_{\mu\nu}$ is the density matrix, and $[\mu\nu|\lambda\sigma]$ are two-electron repulsion integrals over primitive basis functions

* Corresponding author e-mail: todd.martinez@stanford.edu.

$$[\mu\nu|\lambda\sigma] = \iint \chi_\mu(\vec{r}_1)\chi_\nu(\vec{r}_1)\frac{1}{r_{12}}\chi_\lambda(\vec{r}_2)\chi_\sigma(\vec{r}_2)d\vec{r}_1^3 d\vec{r}_2^3 \tag{2}$$

$$\chi_\mu(\vec{r}) = c_\mu(x - X_\mu)^{n_x}(y - Y_\mu)^{n_y}(z - Z_\mu)^{n_z} \times$$
$$\exp(-\alpha_\mu(\vec{r} - \vec{R}_\mu)^2) \tag{3}$$

where $\vec{r} \equiv \{x,y,z\}$ is the electronic coordinate, $\vec{R}_\mu \equiv \{X_\mu,Y_\mu,Z_\mu\}$ is the position of the atomic center associated with the μth basis function, and α_μ determines the width of this basis function. The total angular momentum of χ_μ is given by the sum of the three integer parameters $l = n_x + n_y + n_z$ and is equal to 0, 1, 2, for *s*-, *p*-, *d*-type functions, etc. Our program currently supports only *s*- and *p*-functions, although we are working on implementation of higher orbital momentum functions. The first two sums in eq 1 require little work compared to the last one and are calculated on the CPU in our implementation. All calculations on the CPU are carried out in full double precision, while calculations on the GPU are carried out in single precision unless otherwise indicated. The last sum, which is evaluated on the GPU, combines Coulomb and exchange contributions and runs over only those $\mu\nu$ pairs where at least one of μ or ν is centered on atom A. As in our implementation of direct SCF,[14,15] we treat the Coulomb and exchange terms separately, generating all required two-electron integrals and gradients from scratch using the McMurchie-Davidson algorithm[17]

$$[\mu\nu|\lambda\sigma] = \sum_{pq} E_p^{\mu\nu} E_q^{\lambda\sigma}[\Lambda_p|\Lambda_q] \tag{4}$$

where Λ_p (Λ_q) is a Hermite Gaussian product centered on $\vec{R}_{\mu\nu}^+$ ($\vec{R}_{\lambda\sigma}^+$), and $E_p^{\mu\nu}$ ($E_q^{\lambda\sigma}$) are the expansion coefficients of the Cartesian Gaussian product $\mu\nu$ ($\lambda\sigma$) over the Hermite functions, i.e.

$$\chi_\mu(\vec{r} - \vec{R}_\mu)\chi_\nu(\vec{r} - \vec{R}_\nu) = \sum_p E_p^{\mu\nu}(\vec{R}_{\mu\nu}^-)\Lambda_p(\vec{r} - \vec{R}_{\mu\nu}^+) \tag{5}$$

$$\vec{R}_{\mu\nu}^+ = \frac{\alpha_\mu\vec{R}_\mu + \alpha_\nu\vec{R}_\nu}{\alpha_\mu + \alpha_\nu} \tag{6}$$

$$\vec{R}_{\mu\nu}^- = \vec{R}_\mu - \vec{R}_\nu \tag{7}$$

The nuclear gradients ∇_A and ∇_B can now be represented in the new variables $\vec{R}_{\mu\nu}^+$ and $\vec{R}_{\mu\nu}^-$

$$\nabla_A = \frac{\alpha_\mu}{\alpha_\mu + \alpha_\nu}\nabla_{R^+} + \nabla_{R^-} \tag{8}$$

$$\nabla_B = \nabla_{R^+} - \nabla_A \tag{9}$$

where A and B label the atomic centers of the functions χ_μ and χ_ν, respectively. In the following, we will use the notation $[p|q]$ as a shorthand to indicate the integral over Hermite functions in eq 4, i.e. $[\Lambda_p|\Lambda_q]$. Thus, the indices p and q are pair indices corresponding to $\mu\nu$ or $\lambda\sigma$, respectively. This notation follows that introduced earlier in a comprehensive article by Gill on two-electron integral generation.[18]

A. Coulomb Contribution. The Coulomb contribution to $\nabla_A E$ is generally given by

$$\nabla_A^{Coul}E_{HF} = \sum_{\mu\nu\lambda\sigma} D_{\mu\nu}D_{\lambda\sigma}[\nabla_A(\mu\nu)|\lambda\sigma] \tag{10}$$

Following established practice, we expand the Cartesian Gaussian primitive pair products over Hermite Gaussian basis functions and preprocess the density matrix elements accordingly[19,20]

$$\nabla_A^{Coul}E_{HF} = \nabla_A\left(\sum_p D_p \sum_q D_q[\Lambda_p|\Lambda_q]\right) \tag{11}$$

$$D_p = \sum_{\mu\nu\in p} E_p^{\mu\nu}D_{\mu\nu}, D_q = \sum_{\lambda\sigma\in q} E_q^{\lambda\sigma}D_{\lambda\sigma} \tag{12}$$

Substituting eq 8 into eq 11 leads to the final result for $\nabla_A^{Coul}E_{HF}$

$$\nabla_A^{Coul}E_{HF} = \sum_p \frac{\alpha_{\mu(p)}}{\alpha_{\mu(p)} + \alpha_{\nu(p)}}J_p'D_p + \sum_p J_p\nabla_{R^-}D_p \tag{13}$$

$$J_p = \sum_q D_q[p|q] \tag{14}$$

$$J_p' = \sum_q D_q[\nabla_{R^+}p|q] = -\sum_q D_q[\nabla_r p|q] \tag{15}$$

In the cases under consideration here (*s* and *p* basis functions and no special treatment of *sp*-superblocks), there is only one pair of indices $\mu\nu$ for a given p, and we indicate this with the notation $\mu(p)$ in eq 13. This dependence is suppressed for notational convenience in the following. We calculate J_p and J_p' on the GPU, while the preprocessing of eq 12 and the postprocessing of eq 13 are carried out on the CPU.

Evaluation of J_p and J_p' is performed in a way very similar to our GPU implementation of the Coulomb matrix formation algorithm.[15] Here we use the so-called 1T1PI mapping scheme where one GPU thread calculates one primitive integral (or a batch of integrals if higher than *s* angular momentum functions are involved). This has proved to be the best choice if all required quantities are calculated directly from primitive integrals.[13] The fundamental data organization used for calculating J_p and J_p' is represented in Figure 1, using the same notation as in our previous article.[15] The left and upper triangles represent the index-symmetry pruned lists of the ($[p|$, $[\nabla p|$), and $|q]$ pairwise quantities (PQs) with doubled off-diagonal terms. The PQs are first organized into three groups according to the total angular momentum of the pair products, *ss*, *sp*, or *pp*. Furthermore, the $[p|$ and $|q]$ lists within each angular momentum grouping are sorted according to their Schwartz upper bounds

$$[p|_{Schwartz} = |D_{\mu\nu}|^{max}[p|p]^{1/2} \tag{16}$$

$$|q]_{Schwartz} = |D_{\lambda\sigma}|^{max}[q|q]^{1/2} \tag{17}$$

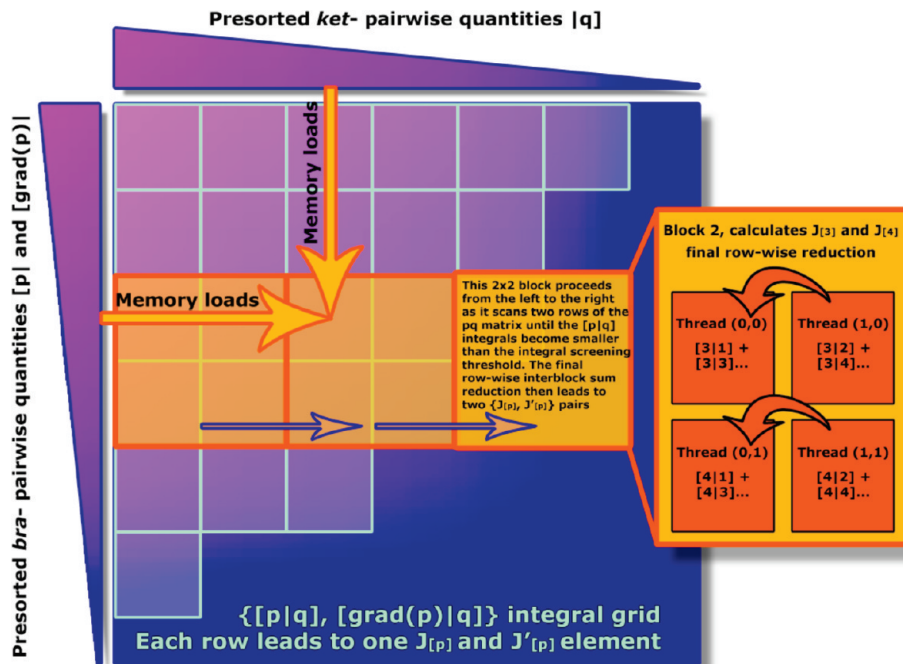where we also use the maximum Cartesian density matrix elements among all angular momentum functions in a batch.

Figure 1. J_p and J'_p calculation algorithm. Each blue-green-bordered square represents four sets of primitive Hermitian integrals ([$p|q$] and [$\nabla p|q$]), which need to be contracted with density matrix elements along a p-row in order to obtain J_p and J'_p. The upper and left triangles denote the *bra*- and *ket*-pair quantities, sorted according to their Schwartz upper bound (represented by the pink-blue coloration). A GPU 2×2 thread block is delineated by a large orange-bordered square comprised of four smaller yellow-bordered squares (GPU threads). The yellow arrows labeled "memory loads" represent elements of the [$p|$ and $|q$] linear arrays loaded into local memory by the corresponding GPU threads. The blue arrows show the direction along which the block scans the two neighboring rows, leading to two $\{J_p, J'_p\}$ pairs of elements. The scan is aborted once each thread in a block encounters a [$p|q$] integral whose Schwartz upper bound is smaller than 10^{-11} au.

Following the usual practice,[19] we calculate the Schwartz upper bound in eqs 16 and 17 assuming p and q have zero angular momentum. The grouping and sorting of the lists of PQs as well as the calculation of the relevant quantities is carried out on the CPU.

When combined, ([$p|$, [$\nabla p|$), and $|q$] lead to the ([$p|q$], [$\nabla p|q$]) integrals required by eqs 14 and 15. These are depicted by blue-green bordered squares in Figure 1 with one square denoting all ([$p|q$], [$\nabla p|q$]) integrals computed simultaneously by one GPU thread. For example, there are $4 \times N_q$ such integrals ([$\Lambda_0|q$], [$\Lambda_x|q$], [$\Lambda_y|q$], and [$\Lambda_z|q$]) if both μ and ν functions have zero angular momentum, and the $|q$] batch has N_q PQs ($N_q = 1, 4, 10$ for *ss*-, *sp*-, *pp*-type batches, respectively). Because of the grouping by angular momentum, the resulting [$p|q$] integral grid consists of nine rectangular segments of different integral types ([$ss|ss$], [$ss|sp$], and so on, up to [$pp|pp$]). Each of these integral grids is handled by a different GPU kernel, optimized for the specific angular momentum. For simplicity, only one such grid segment is presented in Figure 1. The pink-blue coloration in Figure 1 represents the density-weighted magnitude of the Schwartz bounds for the pair-quantities and the resulting [$p|q$] integrals.

It can be easily seen from eqs 14 and 15 that J_p and J'_p can be computed by summing all integral contributions in one row p. Because we use the 1T1PI mapping, each blue-green bordered square in Figure 1 is mapped to a GPU thread depicted as a small orange square. For simplicity, four such GPU threads are depicted, organized into a 2×2 thread block. In practice, we use 8×8 thread blocks in the

implementation. One such block then processes two rows of the integral matrix in column-by-column fashion, accumulating partial results in the corresponding GPU threads with double precision accuracy. Thus, the integrals and integral derivatives are computed in single precision, but the accumulation is done in double precision (on the GPU). As shown previously,[15] this procedure avoids unnecessary precision loss with minimal cost. After the block reaches integrals with a Schwartz upper bound smaller than 10^{-11} au, the scan is aborted, and subsequent intrablock row-wise sum reduction leads to two J_p and J'_p elements. The PQ presorting step guarantees that integrals omitted during the scan are even smaller than 10^{-11} au and thus can be safely disregarded. Because J_p is exactly the same quantity calculated in our previously described Coulomb matrix formation algorithm,[15] while J'_p merely adds some additional terms to be calculated, we simply modified the J-matrix GPU kernels to incorporate these additional terms.

B. Exchange Contribution. The exchange contribution

$$\nabla_A^{Exch} E_{HF} = -\frac{1}{2} \sum_{\mu\nu\lambda\sigma} D_{\mu\lambda} D_{\nu\sigma} [\nabla_A(\mu\nu)|\lambda\sigma] \qquad (18)$$

does not allow easy splitting of the work into the $D_p[p|$ and $|q]D_q$ product representation and thus requires different data organization. Therefore, we generate all required PQs and integrals from scratch without reusing any data from the Coulomb step. In addition, we do not preprocess the density matrix elements on CPU. Instead, the density matrix is preprocessed "on the fly" for each integral

$$\nabla_A^{Exch} E_{HF} = -\frac{1}{2} \nabla_A \left(\sum_p E_p^{\mu\nu} \sum_q D'_{pq} [\Lambda_p | \Lambda_q] \right) \quad (19)$$

$$D'_{pq} = E_q^{\lambda\sigma} D_{\mu\lambda} D_{\nu\sigma} \quad (20)$$

Substituting eq 8 into eq 19 leads to the final result for the exchange contribution to the energy gradient

$$\nabla_A^{Exch} E_{HF} = -\frac{1}{2} \left(\sum_p \frac{\alpha_\mu}{\alpha_\mu + \alpha_\nu} K'_p E_p^{\mu\nu} + \sum_p K_p \nabla_R E_p^{\mu\nu} \right) \quad (21)$$

$$K_p = \sum_q D'_{pq} [p|q], \, K'_p = -\sum_q D'_{pq} [\nabla_r p|q] \quad (22)$$

Unlike in the Coulomb energy gradient calculation, here only the $[\mu\nu|$ pair list is pruned according to $\mu\nu \leftrightarrow \nu\mu$ symmetry (doubling off-diagonal pairs as appropriate), while the $|\lambda\sigma]$ list contains all $O(N^2)$ pairs, where N is the total number of primitive Gaussian-type basis functions. This organization is dictated by the need to carry out prescreening of small integrals in a way that is commensurate with memory access and load balancing requirements on the GPU. Because the $|\lambda\sigma]$ list is not pruned by index-symmetry, there are four rather than three different angular momentum type $\lambda\sigma$ pairs – *ss*, *sp*, *ps*, and *pp*. Thus, the final integral grid contains twelve segments of different integral types (and requires twelve GPU kernels to handle them). Figure 2 provides more details on the GPU implementation of the exchange contribution to energy gradients. For simplicity, only one segment, e.g. *ss|ss*, is represented. Other segments are treated in the same way. We group $[\mu\nu|$ and $|\lambda\sigma]$ pairs according to the first index (μ or λ, respectively). This procedure leads to N blocks, each with fixed $[\mu_i \ldots |$ and $|\lambda_i \ldots]$. Because the $|\lambda\sigma]$ list is not index-symmetry pruned, all $|\lambda_i \ldots]$ blocks contain N Gaussian pairs. The number of pairs in the $[\mu_i \ldots |$ blocks varies (with a maximum of N) because the $[\mu\nu|$ list *is* index-symmetry pruned. All $[\mu_i \ldots |$ and $|\lambda_i \ldots]$ blocks are sorted according to the Schwartz upper bound of corresponding $[\mu_i\nu|$ and $|\lambda_i\sigma]$ pairs. Note that no density matrix information is used in this sorting step. These presorted $[\mu_i\nu|$ and $|\lambda_i\sigma]$ PQs are delineated by left and upper triangles in Figure 2, where the pink-blue coloration represents the Schwartz upper bound magnitude (as in Figure 1). After the PQs are sorted, K_p and K'_p are calculated on the GPU by a series of twelve subsequent GPU kernel calls (one for each angular momentum segment). Figure 2 provides details on the GPU implementation. Here, a 2×2 GPU thread block (we use 8×8 blocks in our program) is tasked to calculate two K_p and K'_p quantities by scanning two nearby rows of the integral grid and accumulating (in double precision, as in the Coulomb algorithm) the partial results in the GPU registers. During the scan, each thread monitors the product of Schwartz integral upper bound and the maximum Cartesian density matrix elements among all angular momentum functions in a batch, i.e.

$$([p|p][q|q])^{1/2} |D_{\mu\nu}|^{max} \quad (23)$$

When this product becomes smaller than 10^{-11} au, a GPU thread aborts processing of the $|\lambda_i \ldots]$ row and resumes from

segment $|\lambda_{i+1} \ldots]$. The fact that $[p|$ and $|q]$ pairs are organized into $|\mu_i \ldots]$ and $|\lambda_j \ldots]$ blocks guarantees that for all threads in a GPU block the maximum density matrix element $|D_{\mu_i\lambda_j}|^{max}$ is the same. Thus, all threads in a GPU block will abort the scan of the $|\lambda_i \ldots]$ segment simultaneously, avoiding potential problems due to load misbalancing. After the scan is complete, the final result is obtained by intrablock sum reduction.

C. Multi-GPU Parallelization. The energy gradient code is parallelized using POSIX threads in order to use all available GPUs in a workstation. For both Coulomb and exchange contributions, the work is split into 8-row segments and mapped to the devices cyclically. Each GPU thus computes its portion of J_p, J'_p, K_p, and K'_p and sends the results back to the host CPU, where they are postprocessed and the final energy gradient is calculated. This model also seems well suited for implementation on a multiple GPU node cluster (using the MPI framework,[21] for example) since the work is distributed in such a way that each node has all the data required to calculate its portion of the integrals from the very beginning, avoiding expensive internode communication. Preliminary implementations support this conjecture, and work along these lines is in progress.

A series of tests performed on a system with two GeForce 295GTX cards, each having two GPU processors, demonstrated reasonable speedup of $3.0-3.5\times$, relative to a single GPU processor. Reported timings include the time required to calculate and sort the pair-quantities, which is currently performed on a single CPU core, as well as the data transfer time required to copy the PQs to the GPU and then copy the results back to the CPU.

Results and Discussion

To assess the performance of the GPU code, we carried out a series of benchmarks on a representative set of test molecules and compared the results to GAMESS[22] ver. 11 Apr 2008 (R1). The GAMESS code was executed on an Intel Core2 quad-core 2.66 GHz CPU with 8GB main memory, which represents a state-of-the-art desktop computing system. All four CPU cores were used in parallel in order to obtain maximum CPU performance. The GAMESS program was compiled with the GNU Fortran compiler and linked with Intel MKL ver. 10.0.3. Our GPU code ran on the same workstation with two Nvidia GeForce 295GTX cards operating in parallel. All performance results in this article correspond to this "workstation to workstation" comparison rather than "a single GPU to a single CPU core" comparison. This provides a realistic assessment of the real performance gain one can obtain from a GPU system. For brevity, this quad-core CPU machine and the dual-GPU machine are referred to as "CPU" and "GPU", respectively.

Table 1 presents the time required to calculate the Hartree−Fock energy gradient vector for caffeine ($C_8N_4H_{10}O_2$), cholesterol ($C_{27}H_{46}O$), buckyball (C_{60}), taxol ($C_{45}NH_{49}O_{15}$), valinomycin ($C_{54}N_6H_{90}O_{18}$), and olestra ($C_{156}H_{278}O_{19}$) molecules using the 3-21G basis set. Among these test systems, the largest has 2131 basis functions. One can see that even for a small molecule such as caffeine, the GPU outperforms the CPU by a factor of
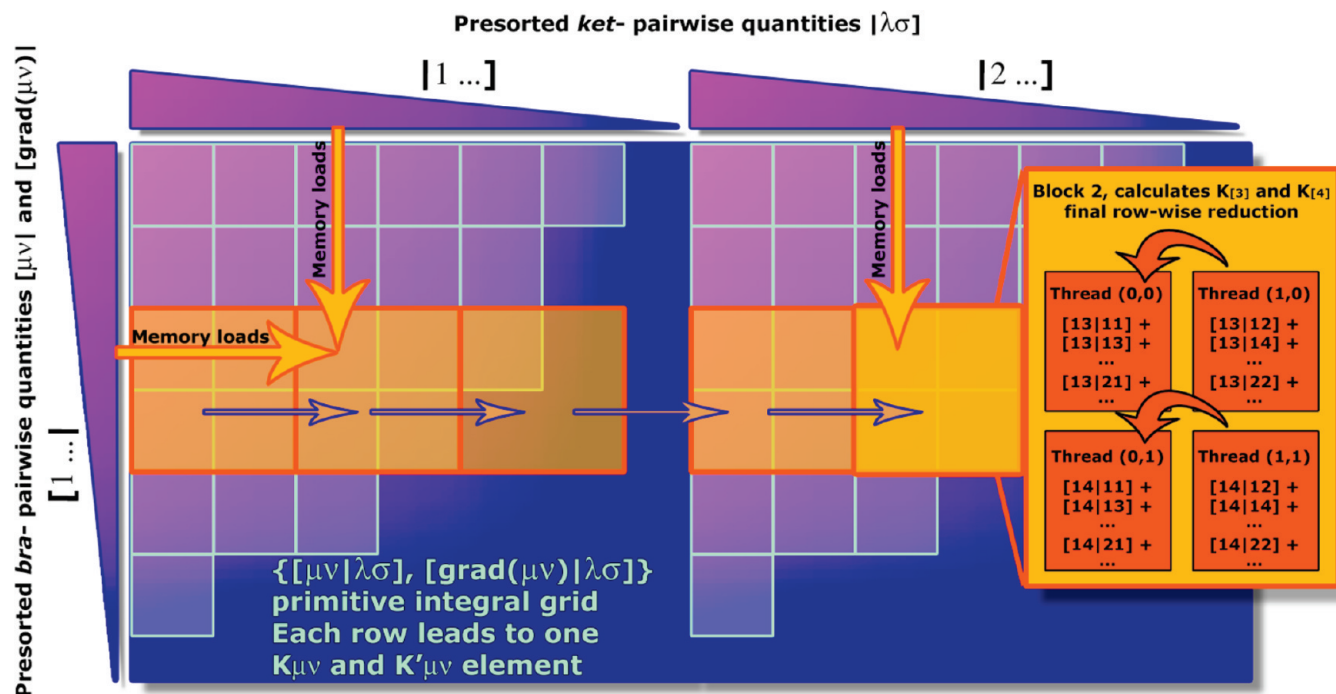
Figure 2. $K_{\mu\nu}$ and $K'_{\mu\nu}$ calculation algorithm. Similar to the *J* calculation shown in Figure 1, but in this case all *ket*-pairs are grouped into *N* segments according to the λ_i index. Two such segments are represented. The GPU thread block scans all the segments sequentially. Once integrals which make contributions smaller than 10^{-11} au are reached, the scan of the particular segment is aborted, and the block proceeds from the next segment.

Table 1. Analytical Energy Gradient Accuracy and Calculation Time for Different Test Molecules Using the 3-21G Basis Set

molecule	caffeine	cholesterol	C_{60}	taxol	valinomycin	olestra
CPU, sec[a]	0.7	7.6	38.5	28.9	66.3	785.1
GPU, sec[b]	0.1	0.4	1.9	1.4	2.6	7.3
Speedup	5.8	19	20	21	26	108
$rms_{error}/$ 10^{-5} au[c]	1.28	0.61	2.79	1.49	1.24	0.67

[a] GAMESS on Intel Core2 quad-core 2.66 GHz CPU. [b] 2 Nvidia GeForce 295GTX cards. [c] Error in the gradient (atomic units) as defined by eq 24.

Table 2. Total Energy and Gradient Computation Time Using the 3-21G Basis Set[a]

molecule	caffeine	cholesterol	buckyball	taxol	valinomycin	olestra
CPU, sec[b]	7.4	81.8	364	435	1112	22863
GPU, sec[b]	2.0	4.8	11.6	15.7	25.5	125
Iterations	13	11	10	14	13	14
Speedup	3.7	17	31	28	44	182

[a] The number of SCF iterations performed to converge the wave function was exactly the same for GPU and CPU. [b] CPU and GPU are the same machines as in Table 1.

6×. For medium-size molecules the speedup ranges between 20× and 25×, while for large molecules it exceeds 100×. In addition, in Table 2 we present the total program execution time, energy + gradient, for the same set of molecules. In both CPU and GPU calculations the total number of SCF iterations required to converge the wave function was exactly the same, although it differed among the molecules. The resulting speedups range from 4× for small-, 30×−40× for medium-, and up to almost 200× for the largest molecules. Clearly, these results represent a mixture of different coding styles, compiler

efficiencies, and hardware architectures. However, it is also obvious that such speedups would never be achievable without the impressive performance gain provided by the GPU, which enables desktop calculation of *ab initio* geometry optimization and molecular dynamics (MD) simulations that were previously only possible on computing clusters with more than a hundred CPUs.

The increased performance of the GPU code clearly improves the quality of molecular dynamics simulation results by allowing longer runs and thus better statistics. However, accuracy is another aspect that needs to be considered, especially when part of the energy and atomic force calculations are performed with single precision. To assess the error introduced by the use of single precision for integral evaluation, we performed three independent tests.

First, we directly calculated the root mean squared error for all components of corresponding single precision (GPU) and double precision (CPU) gradient vectors

$$RMS_{error} = \sqrt{\sum_{i=1}^{3N_{Atoms}} (f_i^{CPU} - f_i^{GPU})^2 / 3N_{Atoms}} \quad (24)$$

for all the benchmark molecules. All molecular geometries along with corresponding energies and gradients, calculated on CPU and GPU, are provided in the Supporting Information. The results, presented in Table 1, demonstrate that the mean error is distributed around 10^{-5} au, which is close to the typical convergence thresholds used in geometry optimization algorithms. It is also important that there is no obvious correlation between the mean error and the size of a molecule, i.e. the error does not increase with the number of atoms.
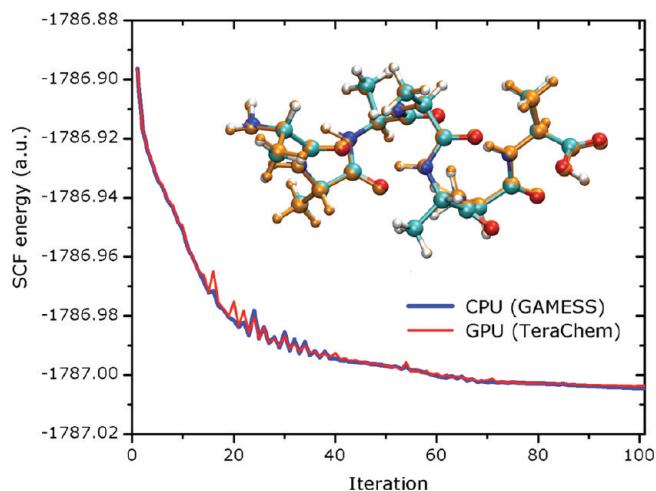
Figure 3. Evolution of the (ALA)$_7$ helix SCF energy during geometry optimization on the CPU (blue) and GPU (red) using 3-21G basis set. Both curves follow each other very closely, demonstrating that the GPU accuracy is sufficient for solving geometry optimization problems. The final energy difference is 0.50 kcal/mol, of which 0.08 kcal/mol is due to lower accuracy of single point energy calculations on GPU and 0.42 kcal/mol is due to different atomic positions. The inset portrays the two optimized structures overlaid on top of each other (orange: CPU, multicolored: GPU).

Figure 4. The SCF (blue), kinetic (red), and total (green) energies of the $H_3O^+(H_2O)_{30}$ cluster during microcanonical (<E_{kin}> 301 K) Born−Oppenheimer Hartree−Fock molecular dynamics simulation using the 3-21G basis set on two Nvidia GeForce 295GTX GPUs. Two electron integrals and their derivatives are calculated on the GPU in single precision, and their contributions are accumulated in double precision. The total energy drift is 0.022 kcal/mol·ps^{-1}, which corresponds to a 0.039 K·ps^{-1} averaged temperature drift.

Second, we carried out geometry optimization of a helical hepta-alanine, i.e. (ALA)$_7$, peptide on the CPU and GPU using the 3-21G basis set. The corresponding SCF energy evolution and resulting structures from GAMESS and our GPU code are compared in Figure 3. In both cases, the same initial geometry and energy minimization algorithm was used (trust region method with BFGS Hessian update), although our implementation was written from scratch and therefore the codes are not necessarily identical. However, we did employ exactly the same trust radius update protocol as implemented in GAMESS. It can be easily seen that the curves follow each other very closely throughout the whole optimization procedure, slightly diverging at the very end. This is a good indicator that GPU can be efficiently used for solving molecular geometry optimization problems. The final energy discrepancy is 0.50 kcal/mol, of which 0.42 kcal/mol is due to different atomic configurations and the rest is due to lower accuracy of single point energy calculations on GPU. In addition, the optimized structures overlap almost perfectly, as shown in the inset of Figure 3, where the CPU-optimized and GPU-optimized structures are portrayed in orange and multicolored representations, respectively.

Finally, conservation of total energy is a common metric for assessing the energy gradient accuracy in a dynamics algorithm. Therefore, we performed time-reversible[23] Hartree−Fock Born−Oppenheimer molecular dynamics simulation of an $H_3O^+(H_2O)_{30}$ cluster using the 6-31G basis set and microcanonical ensemble. The Newtonian equations of motions were integrated using the velocity Verlet algorithm with a 0.5 fs time step for a total simulation time of 20 ps. Figure 4 shows the resulting time evolution of the kinetic (red), potential (blue), and total (green) energies. Energy is conserved quite well, with a small 0.022 kcal/mol·ps^{-1} total energy drift. Consider-
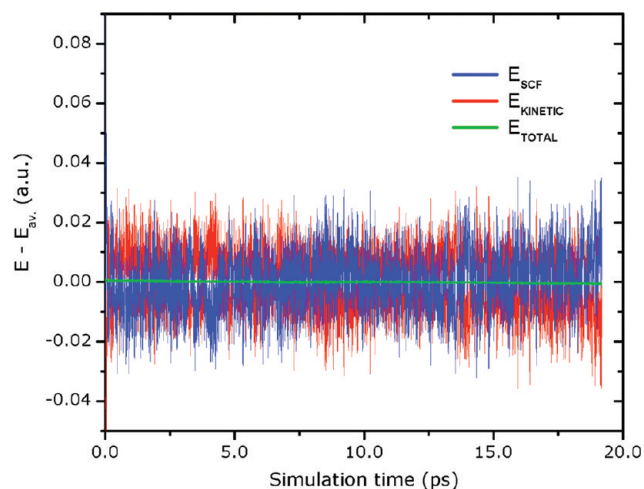
ing, for example, an upper limit of 5% of the initial kinetic energy as a maximum acceptable total energy drift over the entire simulation, this system can be simulated for 200 ps. Even longer time scales can be accessed if one introduces Langevin thermostats and careful adjustment of the damping parameter,[24] although care needs to be taken in this case to ensure that the dynamics is not significantly modified by the damping parameter. To further demonstrate that GPU-based *ab initio* molecular dynamics can treat important phenomena of widespread chemical interest such as proton transfer, we performed AIMD simulations of two systems: a $H_3O^+(H_2O)_{30}$ cluster and a neutral aspartic acid molecule solvated by 147 water molecules.

A. Protonated Water Cluster. A 20-ps long NVT (T = 300 K) molecular dynamics simulation was performed on the $H_3O^+(H_2O)_{30}$ cluster at the RHF/6-31G, BLYP/6-31G, and B3LYP/6-31G levels using a 0.5 fs integration time step. The density functional theory calculations (energy and nuclear energy gradient), which run entirely on GPU and use all available GPU processors in parallel, were recently incorporated into TeraChem, the general purpose GPU-based quantum chemistry package being developed in our group. All three simulations started from an initial cluster geometry where the hydronium ion was located on the cluster surface.

Figure 5 shows the time evolution of the distance between the hydronium ion and the cluster center along with the mean cluster radii represented by dashed lines. The cluster radius was defined as the maximum distance between any of the O-atoms and the cluster center. In all three cases, the ion stayed close to the surface throughout the entire simulation, as previously reported.[25,26] Somewhat larger oscillations of the distance in the RHF
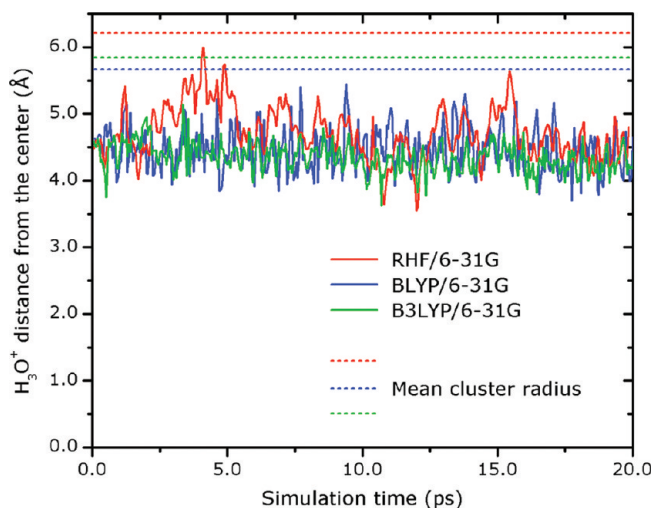
Quantum Chemistry on Graphical Processing Units. 3

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2625**

Figure 5. Time evolution of the H_3O^+ ion distance from the cluster center. Each point on the plot is averaged over 100 successive MD steps. The dashed lines represent the cluster radius, averaged over the whole MD simulation run. The radius is defined as the maximum distance between an O-atom and the center of the cluster.
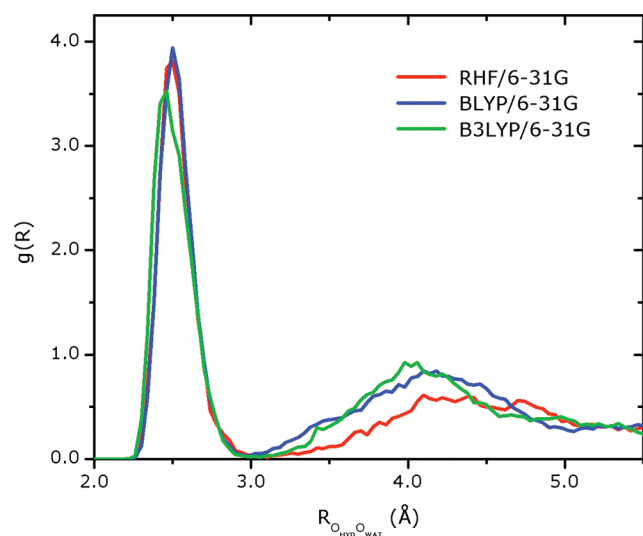

Figure 6. The hydronium-water oxygen−oxygen radial distribution function for the $H_3O^+(H_2O)_{30}$ cluster, using various levels of theory and the 6-31G basis set.

simulation, compared to the DFT results, show that at the HF level of theory the structure of the cluster surface is subject to higher fluctuations. In addition, RHF and BLYP provide similar structures of the first H_3O^+ solvation shell, as can be seen from the oxygen−oxygen radial distribution functions presented in Figure 6 (red and blue lines, respectively). In both cases, the first peak is centered at 2.50 Å, which is smaller than the same value for bulk water (2.8 Å) and is an expected signature of stronger H-bonds. The B3LYP simulation (green line in Figure 6) predicts a somewhat more diffuse first solvation shell, with the first peak centered at 2.46 Å. In addition, the B3LYP RDF reveals stronger bimodal character around the maximum due to continuous interplay between Eigen and Zundel structures. In all three simulations there was an

average of three water molecules in the first solvation shell of the hydronium ion.

B. Solvated Aspartic Acid. We also performed a 7-ps simulation of a single neutral aspartic acid molecule solvated by a cluster of 147 water molecules using the 3-21G basis set. The total number of atoms and basis functions in this system was 457 and 2014, respectively. The $C_4NH_7O_4$ molecule was first solvated in a cube of water and equilibrated for 100 ps using the CHARMM force field[27] and periodic boundary conditions. Then, a 10 Å-sphere was sketched around the C_α atom, and those waters whose O-atoms were located inside this sphere (147 molecules) were selected for further modeling with AIMD. The resulting system was equilibrated for 1 ps using AIMD. In both classical and *ab initio* equilibration MD simulations, a 0.5 fs integration time step and the Langevin thermostat[28] ($T = 300$ K, $\tau_{damp} = 1$ ps) were used, and the atoms in the aspartic acid molecule were fixed in order to prevent deprotonation of the carboxylic groups. Finally, a 7-ps production AIMD run was performed on the system in the NVT ensemble using a 0.5 fs integration time step. Figure 7 portrays a snapshot of the system (left panel) along with its highest occupied molecular orbital (HOMO, right panel). The orbitals were calculated by the GPU-accelerated Orbital plugin implemented in VMD.[29]

The aspartic acid molecule has two carboxylic acid functional groups - backbone and side chain, referred to as COOHbb and COOHsc in the following. Because both the groups have low pK_a, one might expect them to deprotonate quickly in aqueous solution. The AIMD simulation mostly confirms the expectations. Figure 8 displays the time evolution of the O−H distance for both COOH groups. One of these (COOHsc, blue line in Figure 8) quickly deprotonates in approximately 400 fs through formation of a short-lived ($\tau \sim 50$ fs) transient Zundel-like structure. The resulting hydronium ion then quickly (within 50 fs) shuttles to the surface of the cluster via two subsequent proton transfer events and stays at the surface for the rest of the simulation. The second deprotonation event (of the COOHbb group) does not occur until significantly later ($t \approx 5.5$ ps) in the simulation, presumably because of the high proton affinity of the resulting doubly negatively charged amino acid ion. An aborted attempt at deprotonation of COOHbb by forming a quasi-stable $COO^-\cdots H_3O^+$ complex is observed after 3.1 ps of simulation (green line in Figure 8). The presence of positive counterions near the molecule (not accounted for in our simulation) would be expected to facilitate faster deprotonation of all carboxylic groups. The carboxyl-water $O_{carboxyl}$-O_w (including both oxygen atoms of each of the COOHsc and COOHbb groups) and amino-water N_{amino}-O_w radial distribution functions, which are proportional to the local water density around these groups, are presented in Figure 9 and provide details on the solvent structure around these functional groups. The $O_{carboxyl}$-O_w RDF (red line in Figure 9) has its first maximum centered at 2.69 Å, and integration to the first minimum of the RDF reveals that on average 2.5 water molecules are present in the first solvation shell of each carboxyl oxygen. In
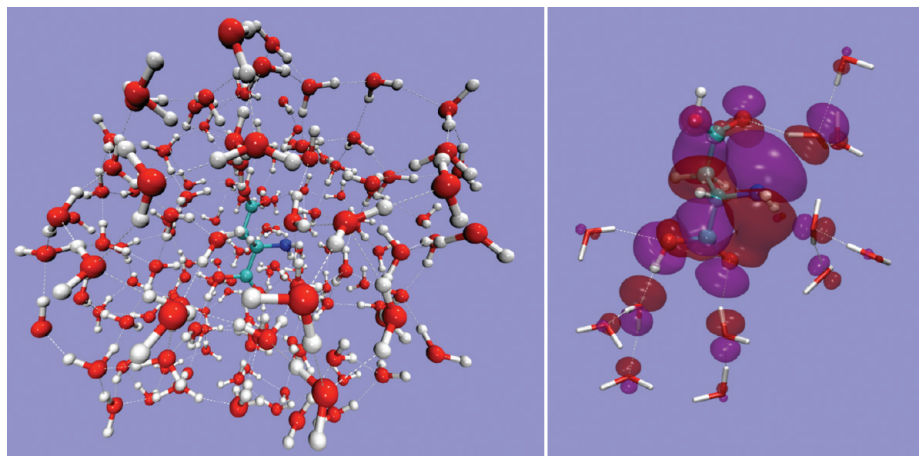
Figure 7. Left: snapshot of the AIMD simulation of the aspartic acid molecule solvated by 147 waters. Right: HOMO along with nearby water molecules among which the orbital is mostly delocalized. The isosurfaces correspond to $\varphi = \pm 0.01$ au .
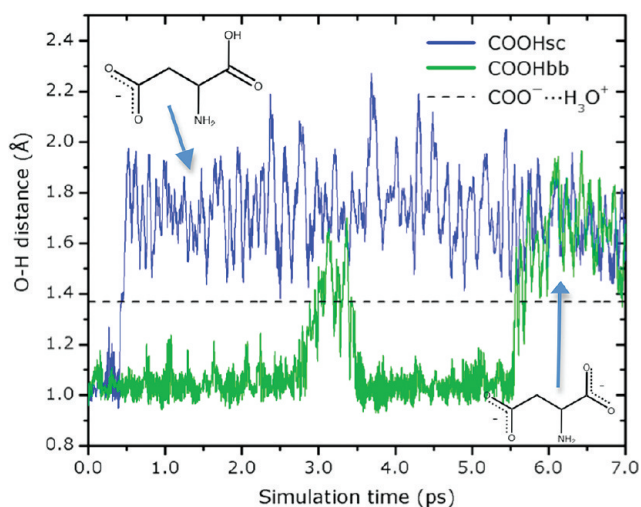
Figure 8. The O−H distance in backbone (COOHbb, green) and side chain (COOHsc, blue) carboxylic acid functional groups. The dashed line represents the O−H distance corresponding to the $COO^- \cdots H_3O^+$ complex.

Figure 9. The aspartic acid oxygen (red) and nitrogen (blue) - water oxygen radial distribution functions. The neutral amino group reveals prominent hydrophobic character.

contrast, the N_{amino}-O_w RDF (blue line in Figure 9) indicates a rather hydrophobic character of the amino group, although the group periodically accepts and donates weak H-bonds. In many cases, however, the solvent water forms a hydrophobic cage around NH_2, where waters prefer to donate/accept H-bonds to/from neighboring waters. Those water molecules that sometimes do accept an H-bond from the amino group, in most cases become 5-coordinated. Because such a water solvation structure is less energetically favorable than the 4-coordinated tetrahedral configuration, these water molecules tend to break such H-bonds. Quantitative analysis of the trajectory demonstrates that the amino group donates 0, 1, and 2 H-bonds for 54%, 45%, and 1% of the simulation time, respectively. Similarly, it accepts 0 and 1 H-bonds for 81% and 19% of the simulation time. An H-bond was defined using 3.2 Å O−O distance and 30° (O/N)$O_w$$H_w$ angle cutoffs. The character of the distribution of H-bonds donated by NH_2 does not vary much during the simulation which accesses three different charge states of the aspartic acid (0, −1, and −2). However, the amino group reveals stronger hydrophilic properties at the −2 charge state by stabilizing the accepted H-bond and reducing the mean N−O_w

distance to the nearest water molecule from 3.1 Å (charge state 0 and −1) to 2.8 Å (charge state −2). Figure 10 shows the time evolution of the N−O_w distance, where the COOH group deprotonation events (for COOHsc and COOHbb) are marked by red vertical lines. The black dashed line denotes the mean N−O_w distance for all (0, −1, and −2) charge states of the aspartic acid molecule before and after the deprotonation of COOHbb. For both neutral and singly negative charge states the distance oscillates around 3.1 Å, but it drops to 2.8 Å after COOHbb is deprotonated and the molecule becomes doubly negatively charged. The hydrophobic-like behavior of the amino group in the partially deprotonated aspartic acid molecule does not seem to preclude protonation. In a preliminary simulation (not shown), the NH_2 residue is rapidly protonated if there is a hydronium ion in its first solvation shell. However, we leave detailed analysis of this for future, more extensive, studies.

Conclusions

We have demonstrated that it is possible to achieve up to $200\times$ speedup in energy + gradient calculations by redesigning quantum chemistry algorithms for the GPU.

Quantum Chemistry on Graphical Processing Units. 3

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2627**
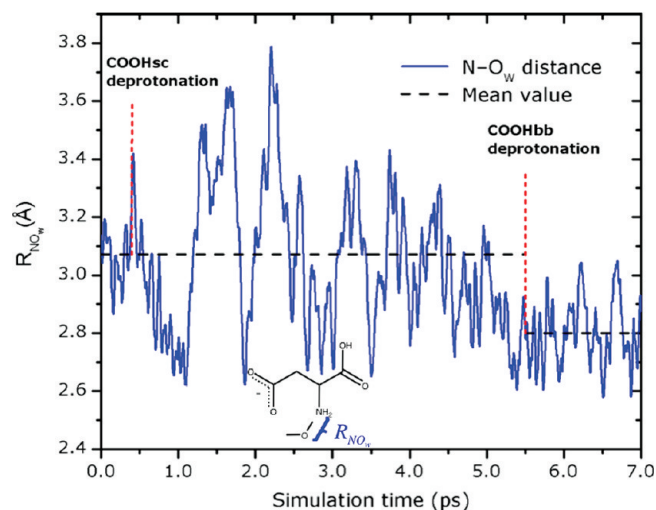
Figure 10. Time evolution of the NO$_w$ distance between the amino nitrogen and the nearest water molecule. Deprotonation events of the carboxylic acid group are marked by red vertical lines. The horizontal dashed lines represent the mean NO$_w$ distance for all (0, −1, and −2) charge states of the aspartic acid molecule. The mean distance is essentially the same (∼3.1 Å) for 0 and −1 charge states but suddenly drops to 2.8 Å as the backbone carboxylic acid group is deprotonated.

The performance gain was assessed by comparing a state-of-the-art Intel Core2 quad-core 2.66 GHz CPU workstation with the same workstation containing two Nvidia GeForce 295GTX graphical cards. Both Hartree−Fock and density functional theory (including generalized gradient and hybrid functionals with exact exchange) electronic structure methods can be used in AIMD simulations with our implementation. The remarkable speedups attained by executing all computationally intensive parts of the code on the GPU rather than on the CPU make it possible to carry out *ab initio* molecular dynamics simulation of large systems containing more than 2000 basis functions at 1400 MD steps/day speed on a single desktop computer.

We show in an example $H_3O^+(H_2O)_{30}$ cluster MD simulation that total energy drift due to limited hardware precision is minor and 100 ps MD simulations can be performed with total energy conservation errors of less than a few percent of the initial kinetic energy. Furthermore, even this minor drift can be compensated by employing a Langevin thermostat and properly adjusting the damping parameter,[24] meaning that even longer *ab initio* MD runs can be performed on the GPU.

We have presented results for preliminary AIMD simulations of proton transfer and transport in solvated clusters that were facilitated by the developments described here. Further study of these phenomena, collecting significant statistics, is underway.

Supporting Information Available: Cartesian coordinates and SCF energies and gradients computed using the CPU (double precision) and GPU (mixed precision) for test molecules listed in Table 1. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lengyel, J.; Reichert, M.; Donald, B. R.; Greenberg, D. P. *Comput. Graph.* **1990**, *24*, 327.

(2) Bohn, C. A. *Joint Conference on Intelligent Systems 1999 (JCIS'98)*; 1998; Vol. 2, p 64.

(3) Hoff, K. E., II; Culver, T.; Keyser, J.; Ming, L.; Manocha, D. *Computer Graphics Proceedings. SIGGRAPH 99*; 1999; p 277.

(4) NVIDIA CUDA. Compute Unified Device Architecture Programming Guide Version 2.2. http://www.nvidia.com/object/cuda_develop.html (acccessed May 15, 2009).

(5) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618.

(6) Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comp. Phys.* **2008**, *227*, 5342.

(7) Liu, W. G.; Schmidt, B.; Voss, G.; Muller-Wittig, W. *Comput. Phys. Commun.* **2008**, *179*, 634.

(8) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864.

(9) Giupponi, G.; Harvey, M. J.; De Fabritiis, G. *Drug Discovery Today* **2008**, *13*, 1052.

(10) Yasuda, K. *J. Comput. Chem.* **2008**, *29*, 334.

(11) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230.

(12) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049.

(13) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222.

(14) Ufimtsev, I. S.; Martinez, T. J. *Comput. Sci. Eng.* **2008**, *10*, 26.

(15) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004.

(16) Pulay, P. *Mol. Phys.* **1969**, *17*, 197.

(17) McMurchie, L. E.; Davidson, E. R. *J. Comp. Phys.* **1978**, *26*, 218.

(18) Gill, P. M. W. *Adv. Quantum Chem.* **1994**, *25*, 141.

(19) Almlof, J.; Faegri, K.; Korsell, K. *J. Comput. Chem.* **1982**, *3*, 385.

(20) Ahmadi, G. R.; Almlof, J. *Chem. Phys. Lett.* **1995**, *246*, 364.

(21) Gropp, W.; Lusk, E.; Skjellum, A. *Using MPI: Portable Parallel Programming with the Message Passing Interface*, 2nd ed.; MIT Press: Cambridge, 1999.

(22) 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.

(23) Niklasson, A. M. N.; Tymczak, C. J.; Challacombe, M. *Phys. Rev. Lett.* **2006**, *97*, 4.

(24) Kuhne, T. D.; Krack, M.; Parrinello, M. *J. Chem. Theory Comput.* **2009**, *5*, 235.

(25) Buch, V.; Milet, A.; Vacha, R.; Jungwirth, P.; Devlin, J. P. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 7342.

(26) Petersen, M. K.; Iyengar, S. S.; Day, T. J. F.; Voth, G. A. *J. Phys. Chem. B* **2004**, *108*, 14804.

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

(28) Schlick, T. *Molecular Modeling and Simulation*; Springer: New York, 2002.

(29) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33.

CT9003004

JCTC Journal of Chemical Theory and Computation

A Hierarchical Approach to Study the Thermal Behavior of Protonated Water Clusters H⁺(H₂O)ₙ

Quoc Chinh Nguyen,[†] Yew-Soon Ong,[‡] and Jer-Lai Kuo[†,§,*]

School of Physical and Mathematical Sciences, Nanyang Technological University, 637371, Singapore, School of Computer Engineering, Nanyang Technological University, 639798, Singapore, and Institute of Atomic and Molecular Sciences, Academia Sinica, Taipei 10617, Taiwan

Abstract: The energy landscape of protonated water clusters H⁺(H₂O)ₙ is thoroughly explored at the first-principle level using a hierarchical search methodology. In particular, the distinct configurational isomers of OSS2 empirical potential for $n = 5-9$ are uncovered and archived systematically using an asynchronous genetic algorithm and are subsequently refined with first-principle calculations. Using the OSS2 model, quantitative agreements in the thermal properties between Monte Carlo and harmonic superposition approximation (HSA) highlighted the reliability of the latter approach for the study of small- to medium-sized protonated water clusters. From the large sets of collected isomers, finite temperature behavior of the clusters can be efficiently examined at first-principle accuracy with the use of HSA. From the results obtained, evidence of structural changes from single-ring to treelike ($n = 5-7$) and multi-ring to single-ring structures ($n = 7-9$) is observed, as expected for the empirical model. Finally, the relevance of these findings to recent experimental data is discussed.

I. Introduction

Protonated water clusters have attracted much study for a long period of time because of their important roles in ionic media and chemical reactions.[1-16] To date, significant progress has been made in experiments[6,7,9,12,13] and theoretical simulations.[3-5,8,11,14-19] In the latter, one of the core focuses is studying the dynamic structural transitions to reveal the complicated thermal behavior of water clusters. Using Monte Carlo (MC) simulations, Singer and co-workers[8] identified the topological transitions of H⁺(H₂O)₈ and H⁺(H₂O)₁₆ to treelike structures at high temperatures, using the OSS2 model.[2] On the other hand, Christie and Jordan,[1] with the use of the MSEVB model,[20] identified two sharp transitions of H⁺(H₂O)₈, whereas that of H⁺(H₂O)₆ was

devoid of sharp structures. Kuo and Klein[4] carried out basin hopping[18,21,22] and a parallel tempering algorithm[23,24] to systematically examine the low-energy structures and structural transitions of H⁺(H₂O)ₙ for *n* up to 21. In a similar work, James and Wales[25] employed a modified empirical valence bond potential to study the properties of selected small- to medium-sized clusters. With the advance of computation power and methodology, several studies have been conducted in an attempt to study water clusters at ab initio levels.[15,16,26-28] Iyengar and co-workers proposed the ab initio atom-centered density matrix propagation method to investigate systematically the structures, dynamics, and vibrational properties of H⁺(H₂O)ₙ for the magic size of *n* = 21,[15,16] as well as other systems.[29] Recently, Nakayama et al.[26] considered the use of an approximate potential to speed up the ab initio MC simulation on small protonated water clusters (H⁺(H₂O)ₙ₌₁,₂).

Most of the existing theoretical simulations have engaged either molecular dynamics (MD) or MC-based algorithms in predicting the structural transitions of H⁺(H₂O)ₙ.[1,4,8,17] Even though some techniques such as J-walk[25,26] and parallel

---

* Corresponding author fax: + 886-223620200, e-mail: jlkuo@pub.iams.sinica.edu.tw.

† School of Physical and Mathematical Sciences, Nanyang Technological University.

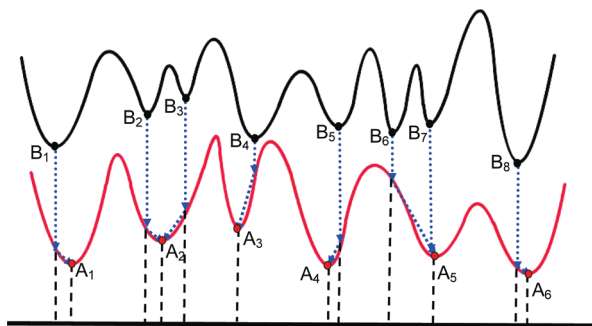‡ School of Computer Engineering, Nanyang Technological University.

§ Academia Sinica.

Figure 1. Sketch of hierarchical approach. A_i and B_i refer to the local minima of the empirical model (black curve) and ab initio calculations (red curve), respectively. The local minima B_i have been identified with the genetic algorithm to serve as appropriate starting points for further refinement or locally optimized using the ab initio calculations to arrive at the respective A_i.
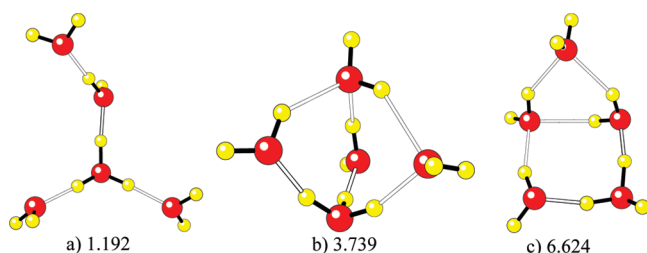

a) 1.192   b) 3.739   c) 6.624

Figure 2. Several new isomers of $H^+(H_2O)_5$. The numbers are denoted as relative energies in kcal/mol.

tempering are established as useful for resolving the issue of quasi-ergodicity, they remain to be plagued with problems of slow convergence, high energy barrier crossing, and poor sampling of the potential energy landscape, especially in large-scale or large-sized systems. An alternative approach to MD and MC is superposition approximation (SA), where the observed quantities of the system are approximated from a statistical set of basins instead of performing sampling and averaging of the entire potential energy surface (PES). To date, SA has been used widely in a variety of systems, from liquids to solids, glasses, clusters, and so forth. It was first considered by Stillinger et al. for studying hidden structures in liquids[30] and subsequently also liquid−solid transitions.[31] Wales examined the coexistence of solid-like and liquid-like forms in a finite atomic cluster, with harmonic approximation employed in SA for the first time.[32] Calvo et al.[33] also used SA to calculate the physical and chemical properties of several atomic clusters and highlighted the close agreement between SA and MC simulations. Generally, SA is established to yield higher efficiency than conventional MC simulations in predicting thermal properties,[33−35] thus making SA feasible and practical for studying the thermodynamics of molecular systems at the quantum chemistry level. Since the accuracy and success of SA depends strongly on how well the PES is explored and how many isomers are taken into account for the approximation, a sampling algorithm for the effective and efficient discovery of true distinct isomers is crucial. Note that this is not a trivial task due to the large numbers of isomers that exist and the high computational costs of ab initio calculations.

Recently, Ohno and co-workers[5,11] proposed the "anharmonic downward distortion following" (ADD) algorithm, a full first-principle-based approach, to explore the PES and then examined the thermodynamics of $H^+(H_2O)_n$ for n up to 7 via harmonic superposition approximation (HSA). The sets of 9, 24, and 131 isomers collected for $n = 5$, 6, and 7, respectively, at B3LYP/6-31+G** level were reasonably large. However, the requirement on second-order derivative calculations and the serial searching regime limits the ADD from sampling widely for high-energy isomers on the quantum chemistry PES, especially for larger-sized systems.

In the present work, instead of a complete reliance on ab initio calculations, we employ a hierarchical methodology with the OSS2 model employed as a prescreening process to construct the archive of potential distinct isomers for $n = 5−9$ to be subsequently examined and refined by first-principles calculations at the B3LYP/6-31+G* level. The synergy between the empirical model and first-principles method permits extensive and efficient exploration of the PES, dealing with large-sized systems (up to $n = 9$), and studying the systems at multiple distinctive levels of theory simultaneously. To deal with the issue of exponential growth of the number of isomers, a parallel asynchronous genetic algorithm is deployed for the searching and archiving of distinct configurational isomers in $H^+(H_2O)_n$. The archive of isomers is then analyzed to investigate the thermodynamic properties and structural transitions of $H^+(H_2O)_n$, $n = 5−10$, for both the OSS2 model and ab initio calculations. From our obtained results, HSA is shown to be reliable in small-sized systems since it reproduces well the feature of heat capacity curves and structural transition of $H^+(H_2O)_n$ in MC simulations of the OSS2 model. The structural transition trends of protonated water clusters in ab initio calculations also revealed close agreement with that of the OSS2 model. The vibrational spectra are derived from the results of HSA and compared with recent experimental study. The details of our computational methodology are described in section II, followed by the results and discussion in section III.

II. Methodology

1. A Hierarchical Approach for Exploration of PES. Our hierarchical approach has been previously explored and applied to the study of pure water clusters.[36] Here, we briefly discuss the main ideas of our methodology as depicted in Figure 1. Instead of searching directly on the PES of ab initio calculations (red curve), which is computationally very expensive, we locate the isomers, denoted here as B_i, of the empirical model (black curve), which serves as a "prescreening" stage. All distinct isomers B_i are subsequently refined to the nearest isomers A_i via ab initio optimizations. The low computational cost of the empirical model allows possible extensive coverage and exploration of the PES for unique isomers. To speed up the entire isomer search process, the ab initio refinements representing the most time-consuming tasks are parallelized on multiple computed clusters.

2. Empirical Models. In this work, the OSS2 model is chosen as the first level of exploration. OSS2 is one of three versions of the OSS family developed by Ojamäe and co-

	W5+	W6+	W7+	W8+	W9+
I	-99.184	-113.854	-129.765	-145.381	-159.635
II	-97.981	-113.834	-129.379	-144.657	-159.127
III	-97.920	-113.602	-129.365	-144.511	-158.400
IV	-96.786	-113.421	-129.244	-144.416	-158.398
V	-96.009	-113.321	-128.235	-144.290	-158.237
VI	-95.445	-113.293	-128.225	-144.116	-158.105
VII	-95.357	-113.247	-128.184	-144.109	-158.021
VIII	-95.275	-113.211	-128.173	-143.730	-157.951
IX	-94.998	-113.132	-128.172	-143.558	-157.920
X	-94.807	-112.977	-128.166	-143.550	-157.769

Figure 3. The 10 lowest-energetic isomers of $H^+(H_2O)_n$, $n = 5-9$, found in B3LYP/6-31+G*, sorted by increasing order of binding energies (in kcal/mol).

workers[2] to simulate water as a participant in ionic chemistry. It was established from the interaction between polarizable O^{2-} anions and protons, with pairwise and three-body terms, and is suitable for studying $H^+(H_2O)_n$ system since it permits the disassociation of water molecules. The potential was parametrized by fitting to ab initio MP2 calculations, and it can reproduce well the structures and binding energies of small-sized protonated water clusters, including neutral ones. For details on the functional forms and thermodynamic simulations of OSS2, the reader is referred to refs 2 and 4.

3. Asynchronous Genetic Algorithm. Archiving isomers is a nontrivial task due to the exponentially increasing number of isomers with increasing cluster sizes. To deal with this issue, we conducted a steady-state genetic algorithm (GA) with an asynchronous scheme for locating isomers, which we have previously used to study the PES of water clusters.[36]

The basic workflow of our GA is outlined as follows: After initialization, parent structures are selected from the population according to their ranks and undergo the genetic operators (crossover and/or mutation) to generate new offspring, which are sent to a slave node for local optimization. The locally optimized structure is then sent back to the master node, which is then updated into the GA population with the spirit of Lamarckian Learning[37,38] and also archived for further analysis. This entire process repeats until a maximum number of iterations is reached. More specifically, the search algorithm is commonly known as a memetic algorithm, hybrid evolutionary algorithm, Lamarckian evolutionary algorithm, cultural algorithm, or genetic local search.[39]

Throughout the archiving process, a significant number of structures, including the low-energetic isomers, may be archived more than once. These duplicate structures not only

	W5+	W6+	W7+	W8+	W9+
I	-89.645	-102.271	-114.160	-125.739	-136.998
II	-89.354	-101.967	-113.689	-125.091	-136.835
III	-89.312	-101.953	-113.657	-125.070	-136.714
IV	-87.334	-101.898	-113.584	-125.047	-136.538
V	-87.123	-101.633	-113.567	-124.979	-136.220
VI	-87.098	-101.606	-113.562	-124.881	-136.216
VII	-87.096	-101.579	-113.518	-124.747	-135.971
VIII	-87.081	-101.569	-113.511	-124.617	-135.943
IX	-86.997	-101.554	-113.397	-124.542	-135.801
X	-86.931	-101.553	-113.365	-124.509	-135.795

Figure 4. The 10 lowest-energetic isomers of $H^+(H_2O)_n$, $n = 5-9$, found in B3LYP/6-31+G*, sorted by increasing order of binding energies with ZPE correction (in kcal/mol).

enlarge the archive unnecessarily but also make the simulation results unreliable since the contribution by some isomers could be artificially amplified. To ensure that only unique structures are archived, the ultrafast structure recognition (USR) algorithm[40] is used to sieve out potential duplicate structures. The similarity of two structures is computed and represented by an index of values ranging from 0 to 1. A value of 0 indicates that two structures are totally dissimilar, whereas the other extreme represent a perfect match. In our study, a newly discovered structure is deemed as a duplicate if the similarity index to existing structures in the archive exceeds a threshold of 0.96.

4. Ab Initio Calculation. All archived distinct isomers of the OSS2 model then undergo geometrical optimization using Becke's three-parameter hybrid method[41] with the Lee, Yang, and Parr (B3LYP) functional[42] and the 6-31+G* basis set. The convergence criteria are defined as a root-mean

square and a maximum component of gradient lower than 0.0003 and 0.00045 hartree/Å, respectively. To ensure that the optimized structures are true isomers, vibration analysis is subsequently performed, and only those having nonimaginary frequencies are accepted for harmonic superposition approximation. Note that all calculations have been completed using the Gaussian 03 package.[43]

III. Results and Discussions

1. Isomers of Protonated Water Clusters. It is worth highlighting that our GA has reproduced successfully all of the most stable structures of $H^+(H_2O)_n$, $n = 5-9$, that were found by the basin-hopping algorithm reported in previous related works.[4] In addition to that, we have uncovered and archived large numbers of distinct isomers which are summarized in Table 1. Since a threshold value of 0.96 was

Figure 5. Canonical heat capacity C_v (upper panel) and population (lower panel) of five topologies for $H^+(H_2O)_n$, $n = 5-10$, that is, multi-ring (MR), double-ring (DR), single-ring (SR), linear (L), and tree-like (T), calculated using C-HSA with the OSS2 model. Solid-red and dashed-blue lines in the upper panels represent heat capacity curves from a previous parallel tempering Monte Carlo (PT-MC) simulation[4] and our C-HSA, respectively. For the sake of brevity, the heat capacity is plotted with a dimensionless unit (Nk_B).

employed in the USR technique to remove duplicate isomers automatically, an overelimination of unique isomers may happen. The resultant archive therefore may not represent the complete set of isomers on the PES. However, from our results, the number of distinct isomers, $N_{isomers}$, is observed to increase exponentially with the number of atoms, N, in the range of $n = 5-9$, as expected in both theoretical and numerical studies.[34,44] $N_{isomers}$ can be approximated as $A \times \exp(\alpha N)$, where both A and α are constants depending on the system considered. From the database of Lennard-Jones clusters, A and α have been approximated as 0.00397 and 0.9897, respectively, for $N \leq 16$.[45,46] In the present work, A and α are approximated as 0.5849 and 0.3918, respectively. The significantly larger A for $H^+(H_2O)_n$ might arise from the fact that the water cluster system is less symmetric than atomic counterparts. This explains the larger number of isomers identified in the small-sized water clusters and, hence, a larger "A" value. Thus, A might represent the measure of asymmetry in the system. In contrast, the smaller value of α of the water cluster system would imply its smaller degree of freedom due to the bonds.

The uncovered sets of archived OSS2 isomers subsequently serve as input structures that are optimized at the

B3LYP/6-31+G* level. At the end of each successful optimization, the local optima at B3LYP/6-31+G* undergo vibrational analysis, which is performed to retrieve the vibrational frequencies. For $H^+(H_2O)_5$, $H^+(H_2O)_6$, and $H^+(H_2O)_7$, a total of 21, 135, and 707 isomers have been uncovered and archived, respectively. For $H^+(H_2O)_8$ and $H^+(H_2O)_9$, because the numbers of isomers found using the OSS2 model are large, a stricter similarity threshold of 0.85 is used to filter out a greater number of potentially duplicated OSS2 isomers before undergoing DFT optimization. On the basis of the given criterion, resultant sets of 422 and 877 isomers for $H^+(H_2O)_8$ and $H^+(H_2O)_9$, respectively, have been identified. In comparison to other recent works reported in the literature from similar studies,[5,11] the archives represent the largest sets of isomers reported at the quantum chemistry level to date. For instance, several isomers of $H^+(H_2O)_5$ that are missing in the recent reported work[11] are depicted in Figure 2.

The 10 lowest-energetic isomers of each cluster size have been sorted according to their binding energies and are depicted in Figure 3. As observed, the ground-state structures of both $H^+(H_2O)_5$ and $H^+(H_2O)_6$ belong to the family of four-membered rings, as observed in the OSS2 model. In
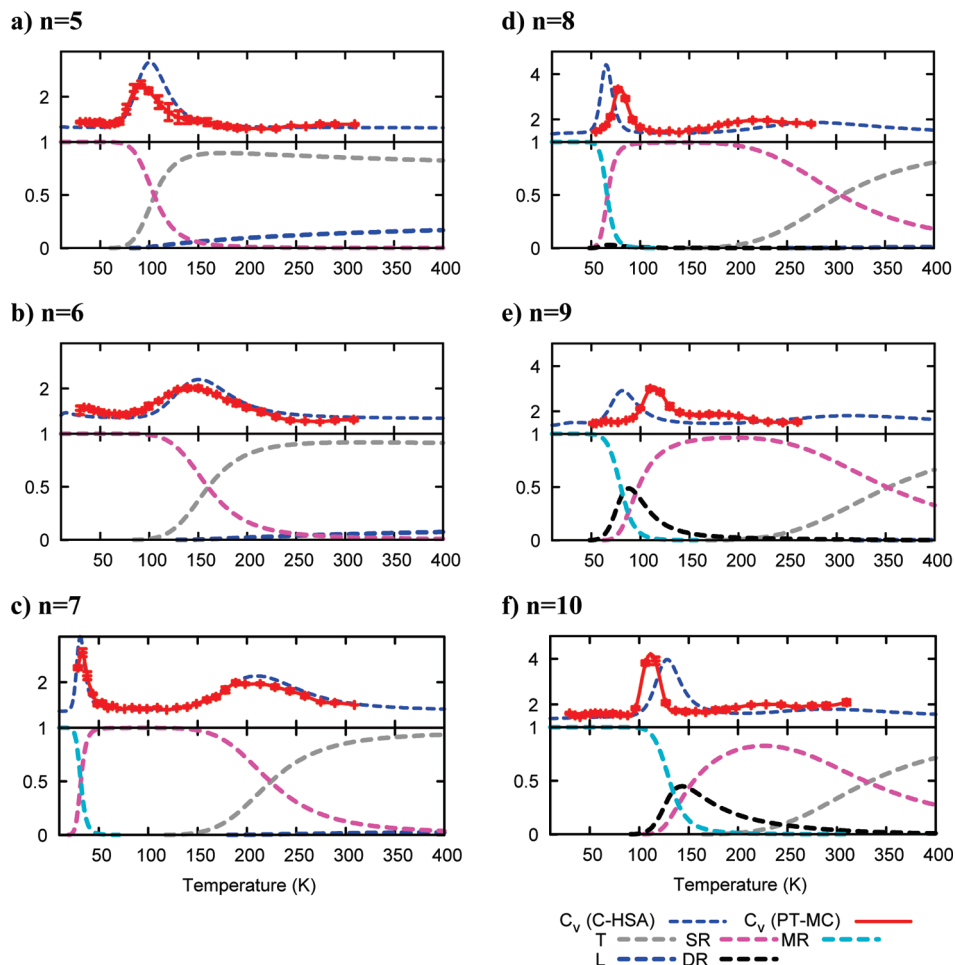
Figure 6. Canonical heat capacity C_v (upper panel) and population (lower panel) of five topologies of $H^+(H_2O)_n$, $n = 5-9$, calculated using C-HSA with B3LYP/6-31+G* calculation. The full details have been described in Figure 5.
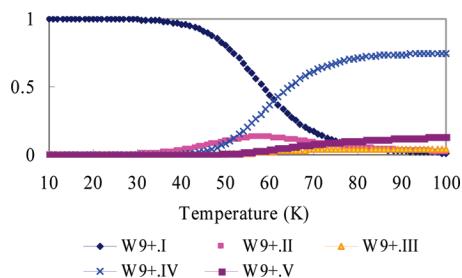
Figure 7. Population plot of the five lowest-energetic isomers of $H^+(H_2O)_9$ calculated using C-HSA with B3LYP/6-31+G* isomers. The transition from the lowest-energetic isomer (W9+).I to the next lowest-energetic ones resulted in the first heat capacity peak of Figure 6e.

$H^+(H_2O)_5$, several other families of isomer shapes are also found to exist, as depicted in Figure 3. These include five-membered ring isomers and tree-like or cage isomers which are well separated at the energy level (~1 kcal/mol). On the other hand, the family of four-membered ring isomers is observed to dominate, with near iso-energetic vakues at a small energy gap of less than 0.2 kcal/mol. For $H^+(H_2O)_7$ and large-sized clusters, the structures tend to be more compact, and all ground-state and low-energetic structures are of multi-ring shapes. The cubic isomers of $H^+(H_2O)_8$

are also relatively stable with a small energy gap of approximately 0.9 kcal/mol to the ground-state counterpart. Note that this is in line with the observation reported by James and Wales[25] where the lowest cubic-like minimum of the modified MSEVB potential is approximately 0.8 kcal/mol higher than the most stable structure.

Figure 4 depicts the structures of $H^+(H_2O)_n$ sorted according to the energy after zero-point energy correction. It is worth noting that the resultant structures of all sizes are observed to be more open. For $H^+(H_2O)_5$ and $H^+(H_2O)_6$, tree-like isomers have the lowest energy instead of single-ring isomers, whereas for $H^+(H_2O)_7$, the single-ring structures dominate. For $H^+(H_2O)_8$ and $H^+(H_2O)_9$, the lowest-energetic isomers are no longer cages, and few single-ring isomers are found. Besides the fact that open isomers are more favored, they are also more iso-energetic since the energy gaps between them are smaller (<1 kcal/mol) than those without zero-point energy.

2. Thermodynamic Transitions. *a. Harmonic Superposition Approximation.* HSA represents an effective approach for acquiring the diverse physical properties of a system from a collection of isomers instead of directly exploring the potential energy landscape. In HSA, each local minimum is treated as a harmonic and infinite basin, which

a) n=5

b) n=6

c) n=7

d) n=8

e) n=9
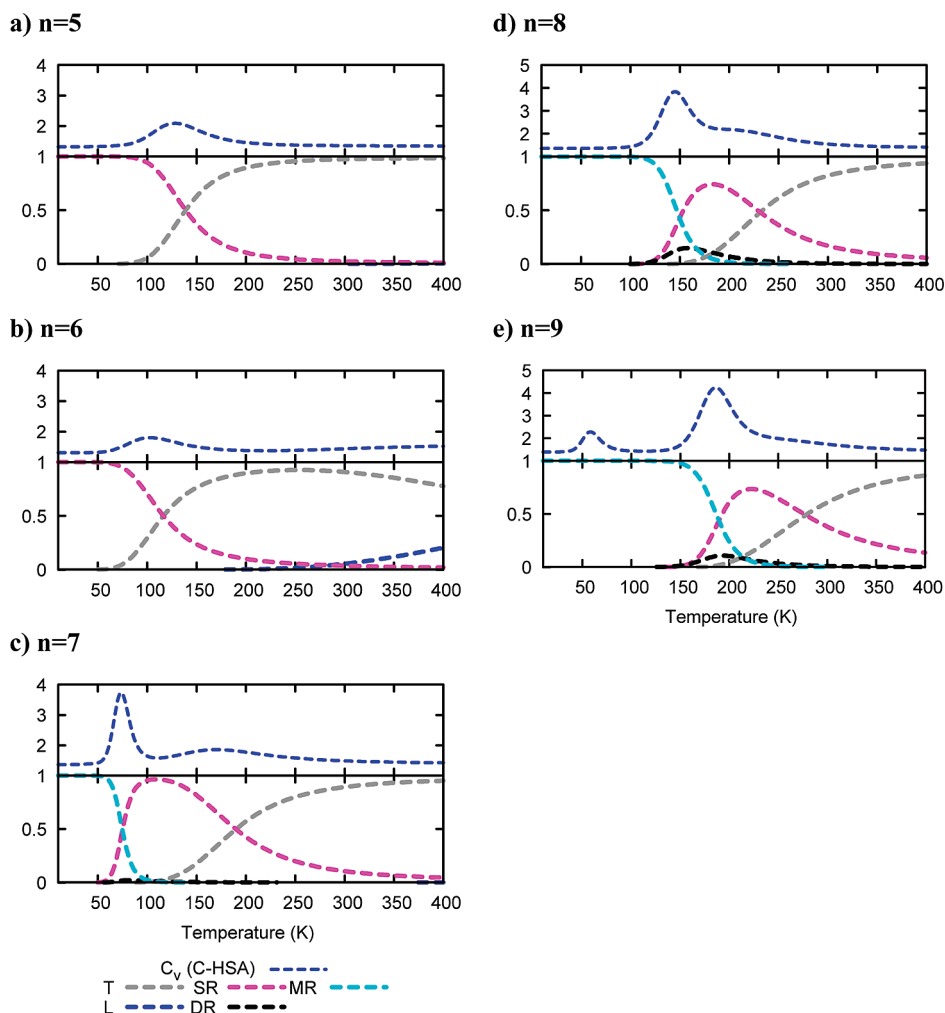
Temperature (K)

C_v (Q-HSA)
T ---- SR ---- MR
L ---- DR ----
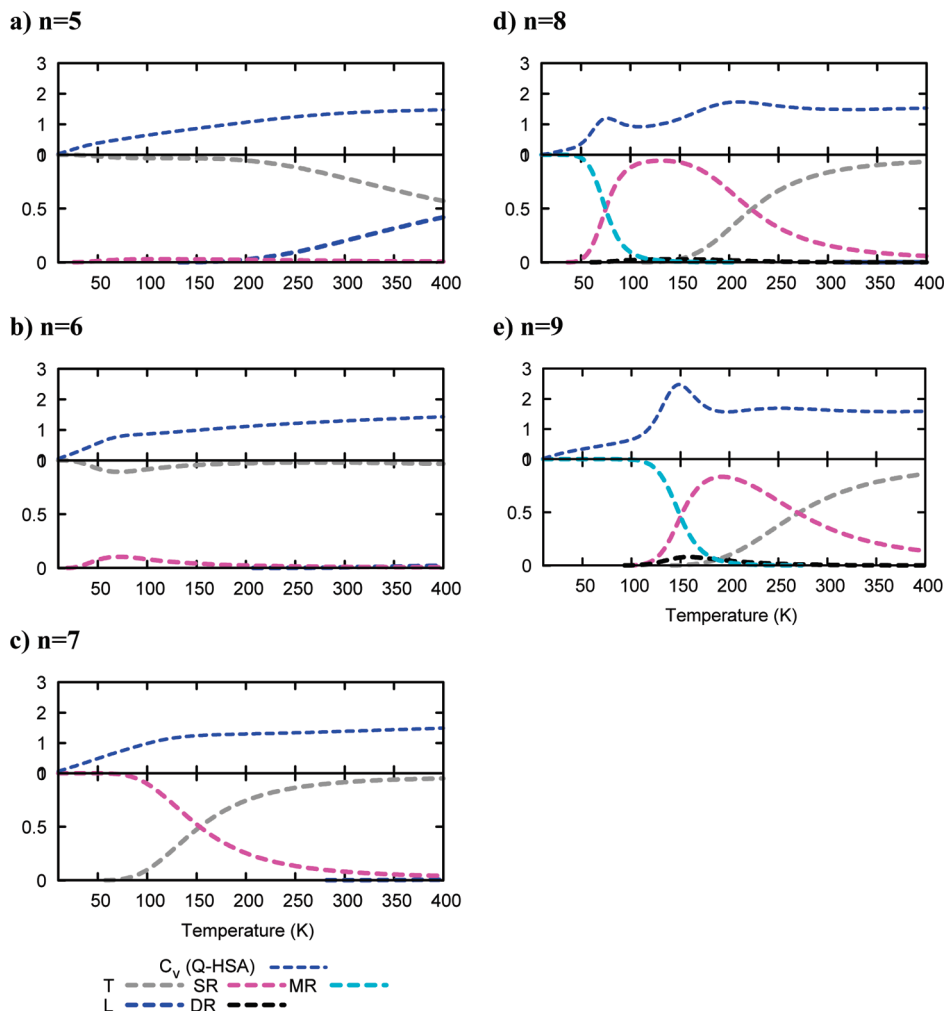
Figure 8. Canonical heat capacity C_v (upper panel) and population (lower panel) of five topologies of $H^+(H_2O)_n$, $n = 5-9$, calculated using quantum HSA (Q-HSA) with B3LYP/6-31+G* calculation. The full details have been described in Figure 5.

is characterized only by its vibrational frequencies and relative energy. From that, observable quantities are subsequently calculated by summing up the contributions of all isomers.

The total partition function $Z(\beta)$ of an N-atom system at temperature T is given by

$$Z(\beta) = \sum_a n_a Z_a(\beta)$$

where $\beta = 1/(k_B T)$ and k_B is the Boltzmann constant. The degeneracy factors n_a in the case of $H^+(H_2O)_n$ (consisting of two kinds of atomic elements, O and H) are derived as $2N_O!N_H!/m_i$, where m_i is the point group order of symmetry. In classical superposition theory (classical harmonic superposition approximation, C-HSA), the harmonic approximation associated with local minima a, $Z_a(\beta)$, is given by[33,34]

$$Z_a^C(\beta) = \exp(-\beta E_a)/(\beta \hbar \overline{\omega}_a)^{N_f}$$

where $\overline{\omega}_a = (\prod_f^{N_f}{}_{=1} \omega_{af})^{1/N_f}$ denotes the mean geometrical vibration frequencies of local minimum a and $N_f = 3N - 6$, representing the number of normal modes. In quantum superposition theory (quantum harmonic superposition approximation, Q-HSA), where the contributions of the vibra-

tion modes are treated in quantum mechanical footing, $Z_a(\beta)$ assumes the following form:

$$Z_a^Q(\beta) = \exp(-\beta E_a) \prod_a \frac{\exp(-\beta \hbar \omega_{af}/2)}{1 - \exp(-\beta \hbar \omega_{af})}$$

To observe the structural transformations, isomers of $H^+(H_2O)_n$ are classified into five categories of topological families, namely, multi-ring, double-ring, single-ring, tree-like, and linear. The population or the canonical probability of the system to fall under topology A is calculated as $P_A(T) = (\sum_{a \in A} Z_a(\beta))/(Z(\beta))$. Other observables including specific heat capacity and free energy can also be easily derived on the basis of the details available in refs 34 and 46.

b. Comparison with Parallel Tempering Monte Carlo (PTMC) Simulation. In Figure 5, the temperature dependences of heat capacity and the populations of five topological families obtained for $H^+(H_2O)_n$, $n = 5-10$, are depicted. For possible comparison, the heat capacity trends obtained by the parallel tempering Monte Carlo (PTMC) simulations as reported in ref 4 are reproduced in Figure 5 with the use of cubic-spline smoothing.

From Figure 5, HSA is observed to reproduce the transition peaks of the heat capacity trace well, on both sizes 5 and 6. Both methods predicted similar transition temperatures in
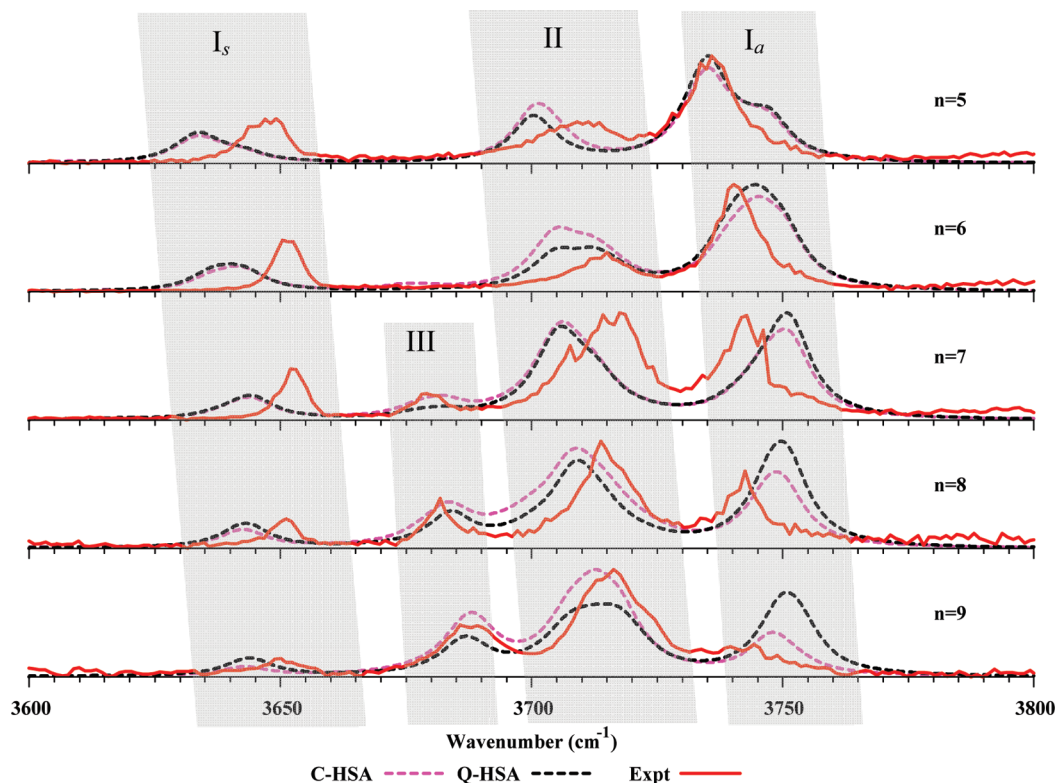
Figure 9. Vibrational spectra of free OH-stretching bands for $H^+(H_2O)_n$, $n = 5-9$, calculated using classical (C-HSA) and quantum (Q-HSA) theories of harmonic superposition approximation. The experimental results of ref 7 are also plotted as a solid-red curve for the purpose of comparison.

Table 1. Number of Distinct Isomers in OSS2 Potential for $H^+(H_2O)_n$, $n = 5-10$

n	OSS2
5	218
6	1192
7	4700
8	11820
9	24693
10	32469

the proximity of 100 K and 150 K for $H^+(H_2O)_5$ and $H^+(H_2O)_6$, respectively. An analysis of the populations of topologies also indicated the domination by single-ring structures dominating at low temperatures and their transformations to tree-like forms. On the other hand, linear structures significantly populating the archive appear only at high temperatures with low probability. A similar observation was also recognized in PTMC simulations.

For $n \geq 7$, where multi-ring structures pose as the most stable, there are two maxima corresponding to two structural transitions: an exceptionally sharp peak for the rapid changes of multi-ring to single-ring structures and a broad and shorter peak for the gradual changes of single-ring to tree-like structures. In the case of $H^+(H_2O)_7$, the transition temperatures predicted by HSA are again in close agreement with those obtained by PTMC; that is, a significant rise at a low temperature of 40 K and a bump at a higher temperature of ~230 K is observed in both. For $H^+(H_2O)_8$, HSA produces slight difference in the prediction of transition temperatures; that is, the first peak (60 K) is lower in PTMC (70 K), as expected, while the second peak shifts to a higher temper-

ature. Nevertheless, the overall feature and trend of the population and heat capacity obtained using HSA correlates well with that using PTMC.

For sizes 9 and 10, the transition temperatures become more difficult to predict accurately since both PTMC and HSA fail to converge well. Note that the increase in transition temperatures due to cluster size in HSA matches that predicted by PTMC previously. These results also revealed transitions from single-ring to tree-like structures happening for all of the cluster sizes investigated and the transition temperature shifting higher with growing cluster size. It is also worth noting the absence of double-ring structures for $n = 5$, 6, 7, and 8 in the population plots, even though double-ring structures significantly populate the archive.

In most cases, except the first peaks of $H^+(H_2O)_8$ and $H^+(H_2O)_9$, HSA seems to have overestimated the transition peak when compared to PTMC, and the shapes of these peaks are noted to be somewhat sharper. However, the overestimation is relatively small and regarded as acceptable for $n \leq 8$, although there is a slight increase for $n \geq 9$. We believe one of the core reasons is related to the underestimations of contributions made by the high-energy local minima, since our exploration of the PES for OSS2 using GA has placed greater emphasis on the low-energy region than the high-energy counterpart. Hence, the entropy of high-energy topologies in the form of tree and linear shapes has generally been underestimated. The resultant impact is a broadening and shifting of the high-temperature transition peak to higher temperatures, while it is a sharpening and swing in the low-temperature transition peak to lower temperatures.

c. DFT Calculation. Employing conventional Monte Carlo methods to simulate thermodynamics in DFT calculations is impractical, even for small-sized systems, due to the high computational cost involved. The use of HSA, on the other hand, serves to be more appropriate due to its higher efficiency compared to conventional MC approaches. Furthermore, HSA has also been shown to yield good quantitative agreement with Monte Carlo simulation on small-sized systems, as demonstrated in the previous sections for the OSS2 model. In what follows, we discuss the use of HSA to study protonated water clusters with DFT calculation in both classical (C-HSA) and quantum (Q-HSA) theory.

i. In Classical Theory. Using the archive of local minima, we applied classical HSA to simulate the thermodynamic transitions. It can be observed in Figure 6 that the heat capacity trend obtained using DFT calculations correlates well with those produced on the basis of the OSS2 model for $n = 5-7$. In the cases of $H^+(H_2O)_5$ and $H^+(H_2O)_6$, a small maximum representing the transition from single-ring to tree-like structures can be observed. The DFT results show a higher transition temperature with a phase change that is more gradual, as reflected by a flat and shorter peak. For $H^+(H_2O)_7$ and $H^+(H_2O)_8$, on the other hand, the two transition points found on DFT are similar to those of OSS2, except with the first occurring at a higher temperature than that in the OSS2 model and the other at a lower temperature. Further, the two peaks of $H^+(H_2O)_8$ are relatively close and appear to merge as single peak in the case of $H^+(H_2O)_9$. Note also the small peak in the heat capacity trend of $H^+(H_2O)_9$ at ~60 K, which corresponds to the transition of the lowest-energetic isomer (denoted as W9+.I in Figure 3) to the second (W9+.II), third (W9+.III), and fourth (W9+.IV) lowest energetic isomers, as depicted in Figure 7. Double-ring isomers appear to start from $H^+(H_2O)_9$ but only survive within a small range of temperature (around 150 K). The overall trend of the structural transitions agrees very well with the OSS2 model, except on the precise positions of the transition temperatures. The size dependency of structural transitions is also consistent with the OSS2 model. The shift to higher temperatures from single-ring to tree-like structures with increasing size is also observed for $T_c = 120, 135, 170, 240,$ and 275 K for $H^+(H_2O)_n$, $n = 5, 6, 7, 8,$ and 9, respectively.

ii. In Quantum Theory. The key difference between quantum HSA and its classical counterpart lies in the involvement of zero-point energy. As mentioned in section III.2.c.i, the inclusion of zero-point energy changes the relative stability of the structures, making open structures such as tree-like and linear forms more favorable. In this section, we discuss their impacts on thermodynamic properties.

The population and canonical heat capacity plots of quantum HSA are depicted in Figure 8. For $H^+(H_2O)_5$ and $H^+(H_2O)_6$, tree-like structures predominate the population in the range of $0-400$ K instead of single-ring forms. As a result, there is almost no structural transition, as reflected in the heat capacity plots. For $H^+(H_2O)_7$, only one maximum of the transformation from single-ring to tree-like structures arises at around 150 K, and in contrast to classical HSA, there is zero contribution by multi-ring structures. It is worth

noting that the observations made are consistent with the recent work reported by Luo and Ohno.[11]

For $H^+(H_2O)_8$ and $H^+(H_2O)_9$, the transition characteristics are similar to classical HSA. Two main transitions, namely, from multi-ring to single-ring and from single-ring to tree-like structures, displayed small contributions of double-ring structures at around 150 K. In addition, the transitions seem to occur at lower temperatures than classical HSA. This makes sense since the zero-point energy correction has the effect of decreasing the entropy of compact structures. As explained earlier in section III.2.b, this causes the shift in transitions to lower temperatures.

3. Vibrational Spectra in Free OH-Stretching Region. From the results of vibrational analysis, the IR spectrum $I_a(\omega)$ of each isomer a is approximated using the Lorentz line shape. The total IR spectrum $I_{total}(\omega, T)$ is then calculated as the weighted sum of $I_a(\omega)$ with the canonical probability $p_a(T)$ of each isomer a derived from thermodynamic simulations, which is given by

$$I_{total}(\omega, T) = \sum_a I_a(\omega) p_a(\omega, T)$$

In Figure 9, the IR spectra calculated using both C-HSA and Q-HSA in the free OH-stretching region are plotted together with the experimental results, which are reproduced from the work of Lin and co-workers.[7] For consistency and fair comparison, the spectra are simulated at temperature $T = 170$ K, which lies within the range of cluster temperatures used in Lin and co-workers' work,[7] as deduced from the Arrhenius equation.[47] In addition, several relevant works[5,11] have typically considered a temperature of 170 K in their investigations on protonated water clusters. Note that all calculated frequencies are also scaled by a factor of 0.973, as suggested in a recent related work.[48]

Four groups of peaks in the free-OH stretching region are highlighted in Figure 9, namely, symmetric free-OH, asymmetric free-OH of one-coordinate H_2O, free-OH of three-coordinate H_2O, and free-OH of two-coordinate H_2O, which are denoted as I_s, I_a, II, and III, respectively. As inferred from the figure, the theoretical spectra derived from both C-HSA and Q-HSA also indicated size dependency, as observed in experimental research. The intensities of peaks I_s (around 3650 cm^{-1}) and peaks I_a (around 3750 cm^{-1}) decrease with increasing cluster size. Nevertheless, our results underestimated the relative decrease of peaks I_a, especially for $H^+(H_2O)_9$, with C-HSA showing better approximation accuracy than Q-HSA, even though the prediction is generally higher than in experimental research. The calculated spectra also agree with the experimental observation that peaks II associated to the free-OH stretching band of three-coordinate H_2O appeared at $n = 7$, and the intensity continues to rise consistently when n increases. As mentioned in the study of Lin et al.,[7] this trend can serve as evidence of a structural transition to ring isomers. Note that this agrees with the thermodynamic results discussed in the previous section where the ring isomers started to predominate the population at $n = 7$. The blue-shift of the whole free-OH stretching bands witnessed in the experiment is also observed in our calculated spectra.

2638 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Nguyen et al.

IV. Conclusions

In this work, we have considered a hierarchical approach to thoroughly explore the PES of $H^+(H_2O)_n$, $n = 5-10$, with the OSS2 potential model and $n = 5-9$ at the B3LYP/6-31+G* level. The distinct isomer set uncovered using our hierarchical methodology is the largest-ever archive found to date. The archive is subsequently used for the investigating the thermodynamic and structural transitions of $H^+(H_2O)_n$ at two theoretical levels, namely, the OSS2 model and the B3LYP/6-31+G* level using the harmonic superposition approximation approach. In comparison with the PTMC simulation involving the OSS2 model, our results revealed good quantitative agreement between HSA and PTMC. The consistency with PTMC results in structural transition and features of the capacity curve are good indications of HSA's reliability. Further, the simulations with the ab initio method also revealed the size dependency of $H^+(H_2O)_n$ in both thermal behavior and vibrational spectra. The calculated vibrational spectra in the free-OH stretching band when compared to recent experimental results also arrived at good agreement.

References

(1) Christie, R. A.; Jordan, K. D. *J. Phys. Chem. B* **2002**, *106*, 8376.

(2) Ojamäe, L.; Shavitt, I.; Singer, S. J. *J. Chem. Phys.* **1998**, *109*, 5547.

(3) Svanberg, M.; Pettersson, J. B. C. *J. Phys. Chem. A* **1998**, *102*, 1865.

(4) Kuo, J. L.; Klein, M. L. *J. Chem. Phys.* **2005**, *122*, 1.

(5) Luo, Y.; Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 10732.

(6) Jiang, J. C.; Wang, Y. S.; Chang, H. C.; Lin, S. H.; Lee, Y. T.; Niedner-Schatteburg, G.; Chang, H. C. *J. Am. Chem. Soc.* **2000**, *122*, 1398.

(7) Lin, C. K.; Wu, C. C.; Wang, Y. S.; Lee, Y. T.; Chang, H. C.; Kuo, J. L.; Klein, M. L. *Phys. Chem. Chem. Phys.* **2005**, *7*, 938.

(8) Singer, S. J.; McDonald, S.; Ojamäe, L. *J. Chem. Phys.* **2000**, *112*, 710.

(9) Wu, C. C.; Lin, C. K.; Chang, H. C.; Jiang, J. C.; Kuo, J. L.; Klein, M. L. *J. Chem. Phys.* **2005**, *122*.

(10) Wu, C. C.; Chaudhuri, C.; Jiang, J. C.; Lee, Y. T.; Chang, H. C. *J. Chin. Chem. Soc.* **2002**, *49*, 769.

(11) Luo, Y.; Ohno, S. M. K. *J. Comput. Chem.* **2008**, *30*, 952.

(12) Headrick, J. M.; Diken, E. G.; Walters, R. S.; Hammer, N. I.; Christie, R. A.; Cui, J.; Myshakin, E. M.; Duncan, M. A.; Johnson, M. A.; Jordan, K. D. *Science* **2005**, *308*, 1765.

(13) Miyazaki, M.; Fujii, A.; Ebata, T.; Mikami, N. *Science* **2004**, *304*, 1134.

(14) Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **1995**, *103*, 150.

(15) Iyengar, S. S. *J. Chem. Phys.* **2007**, *126*, 216101.

(16) Iyengar, S. S.; Petersen, M. K.; Day, T. J.; Burnham, C. J.; Teige, V. E.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 084309.

(17) Mella, M.; Kuo, J.-L.; Clary, D. C.; Klein, M. L. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2324.

(18) Hodges, M. P.; Wales, D. J. *Chem. Phys. Lett.* **2000**, *324*, 279.

(19) Wei, D. Q.; Salahub, D. R. *J. Chem. Phys.* **1997**, *106*, 6086.

(20) Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 9361.

(21) Li, Z. Q.; Scheraga, H. A. Monte-Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*; National Academy of Sciences of the United States of America: Washington, DC, 1987.

(22) Wales, D. J.; Hodges, M. P. *Chem. Phys. Lett.* **1998**, *286*, 65.

(23) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57*, 2607.

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

(25) James, T.; Wales, D. J. *J. Chem. Phys.* **2005**, *122*, 134306.

(26) Nakayama, A.; Seki, N.; Taketsugu, T. *J. Chem. Phys.* **2009**, *130*, 024107.

(27) Iyengar, S. S.; Day, T. J. F.; Voth, G. A. *Int. J. Mass Spectrom.* **2005**, *241*, 197.

(28) Termath, V.; Sauer, J. *Mol. Phys.* **1997**, *91*, 963.

(29) Li, X. H.; Teige, V. E.; Iyengar, S. S. *J. Phys. Chem. A* **2007**, *111*, 4815.

(30) Stillinger, F. H.; Weber, T. A. *Phys. Rev. A* **1982**, *25*, 978.

(31) Stillinger, F. H.; Weber, T. A. *Science* **1984**, *225*, 983.

(32) Wales, D. J. *Mol. Phys.* **1993**, *78*, 151.

(33) Calvo, F.; Doye, J. P. K.; Wales, D. J. *Chem. Phys. Lett.* **2002**, *366*, 176.

(34) Wales, D. J. Properties of Landscape. In *Energy Landscape*, 1st ed.; Saykally, R., Zewail, A., King, D., Eds.; Cambridge University Press: Cambridge, U. K., 2003; Vol. 1, p 364.

(35) Sharapov, V. A.; Meluzzi, D.; Mandelshtam, V. A. *Phys. Rev. Lett.* **2007**, 98.

(36) Nguyen, Q. C.; Ong, Y. S.; Soh, H.; Kuo, J. L. *J. Phys. Chem. A* **2008**, *112*, 6257.

(37) Ong, Y. S.; Keane, A. J. *IEEE Trans. Evol. Comput.* **2004**, *8*, 99.

(38) Turner, G. W.; Tedesco, E.; Harris, K. D. M.; Johnston, R. L.; Kariuki, B. M. *Chem. Phys. Lett.* **2000**, *321*, 183.

(39) Ong, Y. S.; Krasnogor, N.; Ishibuchi, H. *IEEE Trans. Syst., Man, Cybernet., Part B: Cybernet.* **2007**, *37*, 2.

(40) Ballester, P. J.; Richards, W. G. *J. Comput. Chem.* **2007**, *28*, 1711.

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

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

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

Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision D.01; Gaussian, Inc.: Pittsburgh, PA, 2004.

(44) Stillinger, F. H. *Phys. Rev. E* **1999**, *59*, 48.

(45) Wales, D. J.; Doye, J. P. K. *J. Chem. Phys.* **2003**, *119*, 12409.

(46) Bogdan, T. V.; Wales, D. J.; Calvo, F. *J. Chem. Phys.* **2006**, *124*, 1.

(47) Lovejoy, E. R.; Bianco, R. *J. Phys. Chem. A* **2000**, *104*, 10280.

(48) Wang, Y.-S.; Chang, H.-C.; Jiang, J.-C.; Lin, S. H.; Lee, Y. T.; Chang, H.-C. *J. Am. Chem. Soc.* **1998**, *120*, 8777.

JCTC Journal of Chemical Theory and Computation

Simulated Tempering Distributed Replica Sampling, Virtual Replica Exchange, and Other Generalized-Ensemble Methods for Conformational Sampling

Sarah Rauscher,[†,‡] Chris Neale,[†,‡] and Régis Pomès*[,†,‡]

Molecular Structure and Function, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G 1X8 and Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, ON, Canada M5S 1A8

Abstract: Generalized-ensemble algorithms in temperature space have become popular tools to enhance conformational sampling in biomolecular simulations. A random walk in temperature leads to a corresponding random walk in potential energy, which can be used to cross over energetic barriers and overcome the problem of quasi-nonergodicity. In this paper, we introduce two novel methods: simulated tempering distributed replica sampling (STDR) and virtual replica exchange (VREX). These methods are designed to address the practical issues inherent in the replica exchange (RE), simulated tempering (ST), and serial replica exchange (SREM) algorithms. RE requires a large, dedicated, and homogeneous cluster of CPUs to function efficiently when applied to complex systems. ST and SREM both have the drawback of requiring extensive initial simulations, possibly adaptive, for the calculation of weight factors or potential energy distribution functions. STDR and VREX alleviate the need for lengthy initial simulations, and for synchronization and extensive communication between replicas. Both methods are therefore suitable for distributed or heterogeneous computing platforms. We perform an objective comparison of all five algorithms in terms of both implementation issues and sampling efficiency. We use disordered peptides in explicit water as test systems, for a total simulation time of over 42 μs. Efficiency is defined in terms of both structural convergence and temperature diffusion, and we show that these definitions of efficiency are in fact correlated. Importantly, we find that ST-based methods exhibit faster temperature diffusion and correspondingly faster convergence of structural properties compared to RE-based methods. Within the RE-based methods, VREX is superior to both SREM and RE. On the basis of our observations, we conclude that ST is ideal for simple systems, while STDR is well-suited for complex systems.

Introduction

Achieving complete (or even adequate) conformational sampling is one of the key challenges in biomolecular simulations.[1] The energy landscape of most biomolecules is "rugged", and the source of this ruggedness is two-fold. The energetic barriers separating accessible states are often larger than the available thermal energy, and there are typically a large number of states to be sampled. The time scales of many biomolecular processes, such as protein folding, are still far beyond the reach of our current computational capability, which is generally limited to the 10^{-8} to 10^{-7} s time scale for continuous simulations. For example, even the folding of small domains or secondary structure elements, such as β-hairpins and mini-proteins, occur on the $1-10$ μs time scale.[1] Consequently, conventional or "brute force" molecular dynamics (MD) alone is often insufficient to

---

* Corresponding author e-mail: pomes@sickkids.ca.
† Hospital for Sick Children.
‡ University of Toronto.

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2641**

achieve complete Boltzmann sampling of the important states of many biologically relevant systems. For this reason, generalized-ensemble algorithms have become popular tools for conformational sampling.

A variety of generalized-ensemble algorithms have been developed with the common intention of overcoming energetic barriers in order to enhance sampling of conformational space. These methods use a generalized Hamiltonian for the purpose of achieving uniform sampling along a reaction coordinate of interest. Practically, one is faced with choosing the most appropriate method and reaction coordinate for a particular application. While the optimal reaction coordinate is not known a priori, it may be possible to make generalizations regarding the optimal methodology. To this end, we consider the following important question: given limited computational resources, which algorithm is most efficient at sampling a complex energy landscape? Some generalized-ensemble methods employ a random walk in potential energy, while others use different parameters which are relevant to the system of interest.[2] In this article, we compare the efficiency of a set of algorithms which make use of a random walk in temperature to enhance conformational sampling of biomolecules. We focus on the following five methods: simulated tempering (ST),[3,4] replica exchange (RE),[5−9] the serial replica exchange method (SREM),[10] and two novel methods, virtual replica exchange (VREX) and simulated tempering distributed replica sampling (STDR), which is a combination of ST and distributed replica sampling (DR).[11−13]

The generalized-ensemble algorithms compared in this paper all rely on the fact that the free energy surface becomes less rugged at high temperatures, increasing the frequency of interconversion between conformational states.[14] Simulations performed at low temperatures often require a relatively long time to cross the energetic barriers between states and appear to be trapped. Transitions between regions separated by barriers may not be observed over time scales accessible to simulation. In this case, multiple simulations initiated in different conformational basins may sample different subsets of phase space. The result is that an ergodic system appears nonergodic, a phenomenon known as quasi-nonergodicity.[15] Utilizing generalized-ensemble algorithms that induce a random walk in temperature may alleviate this source of error.

The sampling enhancement of generalized-ensemble methods relative to canonical MD or Monte Carlo (MC) simulations has been demonstrated for several systems,[3,7,16,17] including peptides.[6,14,18−24] Conversely, there have also been studies that question the relative sampling efficiency of RE compared to brute force MD,[25] highlighting the importance of a rigorous definition of efficiency which accounts for the total computer time required for all temperatures.[26−28] It is important to note that data obtained at multiple temperatures in generalized-ensemble simulations may be of interest in some studies, such as protein folding.[21,22] In general, however, the data at high temperatures are not useful. Furthermore, the observed speedup also strongly depends on the lowest temperature.[26] It is essential to assess the convergence of both the conventional MD simulations as

well as the generalized-ensemble simulations in order to perform a meaningful comparison, in addition to identifying a meaningful quantity on which to base the comparison. Any evaluation of sampling enhancement compared to single-temperature MD is also likely to depend heavily on the molecular system under study (depending on the number of basins in the landscape and the heights of barriers). It is therefore quite difficult to accurately quantify the sampling enhancement due to the introduction of a random walk in temperature.

We begin with a brief introduction of each of the generalized-ensemble methods, including the presentation of our two novel methods, STDR and VREX. We then perform a thorough comparison of the algorithms in terms of both practical implementation limitations and sampling efficiency for a disordered octapeptide in explicit water, a molecular system combining high relevance to protein folding and moderate complexity. In addition to providing a comparison between generalized-ensemble algorithms, we also provide a comparison to conventional MD. We discuss efficiency in terms of both convergence of structural properties and temperature diffusion, and we show that these definitions of efficiency are correlated. Finally, we compare the efficiency of STDR and conventional MD for a 35-residue peptide with a complex conformational landscape.

Theory and Methods

Simulated Tempering (ST). Simulated tempering was originally introduced to enhance sampling of a random field Ising model.[3] This system has a rough energy landscape for which spin-flips from the state favored by the magnetic field to the opposite state are statistically rare events. ST facilitates exchanges between these states, whereas the MC algorithm remains trapped.[3] ST has also been shown to be effective in exploring the energy landscapes of biomolecules, which similarly have multiple energy minima separated by barriers.[29]

In the ST algorithm, temperature becomes a dynamic variable[3,4] that can take on discrete values labeled by an index m ($m = 1, ..., M$). ST makes use of a generalized Hamiltonian, $H(X,m)$, which depends on all configurational degrees of freedom (X), in addition to temperature:

$$H(X, m) = \beta_m H(X) - a_m \tag{1}$$

where β_m is the inverse temperature, $H(X)$ is the system's original Hamiltonian, and a_m is a constant which depends on temperature.[3] The generalized ensemble has a corresponding generalized partition function, $\mathbf{Z}$, given by:

$$\mathbf{Z} = \sum_m \int dX \, [e^{-H(X,m)}] = \sum_m \int dX \, [e^{-\beta_m H(X) + a_m}] =$$

$$\sum_m Z_m \, e^{a_m} \tag{2}$$

where Z_m is the partition function corresponding to the temperature T_m.[30] The partition function of the generalized ensemble, $\mathbf{Z}$, is the weighted sum of the partition functions of the canonical ensembles at each temperature, Z_m. We

therefore refer to the constants, a_m, as "weight factors".[30] The probability of sampling a given temperature, T_m, is[3]

$$P(T_m) \propto e^{-H(X,m)} \equiv Z_m e^{a_m} \qquad (3)$$

which depends on the generalized Hamiltonian, H, and therefore depends on the weight factor, a_m. The goal in ST is to perform a random walk in temperature such that all temperatures are visited uniformly, that is, to choose weight factors such that for any two temperatures (labeled i and j)

$$Z_i e^{a_i} = Z_j e^{a_j} \qquad (4)$$

Since the partition function in the canonical ensemble, Z_m, is related to the Helmholtz free energy, A_m, the optimal weight factors are dimensionless Helmholtz free energies (the Helmholtz free energy multiplied by the inverse temperature, β):[30,31]

$$\begin{aligned} Z_m &= e^{-\beta_m A_m} = e^{-a_m} \\ a_m &= -\ln Z_m \end{aligned} \qquad (5)$$

The use of accurate dimensionless Helmholtz free energies as weight factors leads to sampling all temperatures with equal probability. In principle, the weight factors may take any value without resulting in biased, non-Boltzmann sampling at the individual temperatures. However, inaccuracy in the weight factors leads to corresponding differences in the probabilities of sampling at each temperature.[4,32]

An ST simulation consists of a short canonical MD (or MC) simulation at temperature T_i followed by an exchange attempt to a neighboring temperature, T_j. The probability of this exchange occurring is given by:

$$p(T_i \rightarrow T_j) = \min\begin{cases} 1 \\ e^{-(\beta_j - \beta_i)E + (a_j - a_i)} \end{cases} \qquad (6)$$

where E is the potential energy of the system at the end of the previous simulation at temperature T_i, and β_i and β_j are the inverse temperatures.[2] The weight factors need only be accurate up to an additive constant, since only differences in weight factors are required to determine the acceptance probability.[30] Through many repetitions of these alternating simulation and exchange steps, a random walk in temperature is realized, corresponding to a random walk in potential energy and efficient exploration of the energy landscape.[31] In fact, ST has been shown to be as effective as the multicanonical algorithm, which employs a random walk in potential energy.[33]

The underlying challenge in ST is accurately obtaining the dimensionless Helmholtz free energies, a_m. There have been two general approaches to their calculation. The first method involves making use of the weighted histogram analysis method (WHAM)[34−36] to obtain the density of states and the weight factors. The second method, which we utilize in this paper, was recently proposed as a fast and efficient scheme to obtain an accurate estimate of the weight factors based on average energies.[30,37] The average potential energy at each temperature, $\langle E \rangle$, is obtained from initial simulations, and the differences in weight factors are calculated as follows:

$$a_{i+1} - a_i \approx (\beta_{i+1} - \beta_i)\left(\frac{\langle E \rangle_i + \langle E \rangle_{i+1}}{2}\right) \qquad (7)$$

The weight factor for the lowest temperature can be set to zero since only differences in weight factors are needed in the exchange probability. The replica exchange simulated tempering method (REST) may also be used to obtain weight factors. In this method, an initial RE simulation is run for the purpose of obtaining accurate weight factors, which are then used in a ST simulation.[36,38] REST may be used with either WHAM or the method outlined in eq 7. Weight factors may be updated throughout the ST simulation if required.[14] Adaptive determination of weights using schemes like the one developed by Zhang and Ma[39] based on the Wang−Landau algorithm[40] were shown to be useful for a square lattice Ising model, a bead model of a protein and a Lennard-Jones fluid.[39] Biomolecular systems containing thousands of atoms represent a completely different level of complexity. Convergence of weights is nontrivial[14,30,31,39] and can be slow even with adaptive approaches.[30] Calculating weight factors has been the main impediment to the widespread use of the simulated tempering method.[30]

The accuracy of the weight factors (that is, how close the differences in weight factors are to accurate dimensionless Helmholtz free energy differences) can be assessed by computing the deviation from sampling all temperatures homogeneously in a sufficiently long ST simulation. In the extreme case for which all weight factors are equal and all differences in weight factors are zero, only the lowest temperature is significantly sampled. This is because the first term in the exponent of the exchange probability (eq 6) depends on the potential energy, which is generally a large, negative number for biomolecular systems. When multiplied by the difference in inverse temperatures, the resulting exchange probability dictates that moves to lower temperatures are accepted, while moves to higher temperatures are rejected. Conversely, if the differences in weight factors are equal to the differences in dimensionless Helmholtz free energies, the temperatures in the ST simulation are sampled uniformly, which is the optimal situation. In practice, weight factors obtained for ST result in temperature sampling inhomogeneity somewhere between these two extremes. Calculating the dimensionless Helmholtz free energies for a complex system such as a peptide in explicit water is computationally expensive since it requires an accurate estimate of the partition function. These calculations can require tens of nanoseconds per temperature or more, and the computational expense increases with both system size and complexity.[14]

Replica Exchange (RE). Replica exchange has been the most widely used of the methods we discuss in this paper to enhance sampling of biomolecular simulations. It can be thought of as a parallel version of ST, and it is also known as parallel tempering[5] or multiple Markov chains.[8] In fact, parallel tempering was applied to proteins even before ST.[41] An RE simulation consists of M identical copies of the system (replicas) which sample M canonical ensembles at different temperatures. Exchanges are performed between neighboring temperatures, T_i and T_j. The probability of

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2643**

making an exchange depends on the potential energies, E_i and E_j, and the inverse temperatures, β_i and β_j:

$$P(T_i \leftrightarrow T_j) = \min \left\{ \begin{array}{l} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_j)} \end{array} \right. \quad (8)$$

RE is analogous to ST, but instead of using weight factors in the exchange probability, the upward move of one replica is coupled to the downward move of another. RE therefore has the critical advantage of not requiring any initial simulation for the calculation of weight factors. Importantly, it also satisfies detailed balance.[2]

One drawback of the RE method is its significant computational requirements. There is a one-to-one correspondence between the number of replicas (M) and the number of temperatures (M). The number of replicas needed for a RE simulation is related to the number of degrees of freedom, N, as O($N^{1/2}$).[2,7,42] Systems with many particles therefore require many replicas. Although it is not a specific requirement of the RE algorithm, in its typical implementation, each replica is run on a dedicated central processing unit (CPU). This setup minimizes the amount of information that must be passed between nodes.[31] Thus, M CPUs are running simultaneously throughout the course of the RE simulation. The use of M CPUs in RE can be overcome by running multiple replicas per CPU. However, using one CPU for multiple replicas does not effectively take advantage of the parallelization inherent in the RE method.

The RE algorithm requires the synchronization of attempted moves, which results in wasted CPU time if any replica waits for other replicas to perform exchanges. Inhomogeneity of CPU speeds affects the amount of wasted time, since the speed of the calculation depends on the speed of the slowest processor. Modified versions of RE have been developed in an effort to minimize wasted CPU time, including the multiplexed replica exchange method (MREM)[23] and asynchronous replica exchange.[43] MREM makes use of multiplexed layers of replicas (n layers, each with M temperatures), with exchanges occurring both within and between layers.[23] MREM is even more computationally demanding than RE, using n times as many processors. MREM does not offer a significant advantage if there is a shortage of CPUs, but it does offer a way of using more CPUs without adding more temperatures. In asynchronous replica exchange, only the replicas undergoing exchange are synchronized, therefore increasing efficiency on heterogeneous computing platforms.[43] More complex replica management schemes have also been proposed to increase the efficiency of RE.[44,45] However, modified RE algorithms do not completely alleviate the need for synchronization and frequent communication between replicas.[14] This is especially important to users of distributed computing, such as the massively parallel Folding@Home project,[46] who must contend with inhomogeneity of processor speeds.[14]

Serial Replica Exchange (SREM). The serial replica exchange method[10] was recently developed to address the main practical limitations inherent in the RE method, namely, the need for synchronization and a large number of processors. The exchange probability in SREM has an identical form to that of RE (eq 8) for a replica at temperature T_i attempting to move to a neighboring temperature T_j:

$$P(T_i \rightarrow T_j) = \min \left\{ \begin{array}{l} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_{j,\text{PEDF}})} \end{array} \right. \quad (9)$$

Unlike RE, the attempted move from T_i to T_j does not simultaneously involve another replica moving from T_j to T_i. In SREM, the potential energy, $E_{j,\text{PEDF}}$, does not come from another replica at temperature T_j but rather is selected at random from a potential energy distribution function (PEDF) for that temperature. The PEDFs are determined through initial simulations at each temperature, which may use either constant-temperature MD or RE. These initial simulations can be very computationally demanding for biomolecular systems. For example, to obtain converged PEDFs for a small RNA hairpin, approximately 100 ns per temperature was required.[47] PEDFs may also need to be updated throughout the course of the SREM simulation.[14,47] SREM also cannot be applied to temperature-dependent force fields.[48,49]

In terms of practical implementation, SREM offers the same advantages as ST. In both methods, there is absolutely no communication required between independent simulations. Neither method requires a fixed number of CPUs, and there is no wasted CPU time in the synchronization of attempted exchanges. In principle, both ST and SREM can be run on a single CPU. SREM also presents the same critical challenge as ST: an initial simulation is needed to determine PEDFs, the length of which is highly dependent on system complexity. The significant computational cost of calculating accurate PEDFs is a key drawback of SREM, since an SREM simulation is not strictly correct if unconverged or incorrect PEDFs are used.[10,14] In contrast, the weight factors of ST can deviate from the accurate dimensionless Helmholtz free energies and still yield correct results.[3,10,14]

Virtual Replica Exchange (VREX). The first novel method we propose, virtual replica exchange, is based on the principles of both RE and SREM. A replica at temperature T_i attempts a move to temperature T_j, with the probability of exchange given by the following equation:

$$P(T_i \rightarrow T_j) = \min \left\{ \begin{array}{l} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_{j,\text{virtual}})} \end{array} \right. \quad (10)$$

Here, the potential energy, $E_{j,\text{virtual}}$, comes from a list of stored energy values obtained at temperature T_j. This is analogous to exchanging with a potential energy value selected from a PEDF in SREM, or the current potential energy of a replica at temperature T_j in RE. Like SREM, only a move from temperature T_i to temperature T_j occurs, with no simultaneous reverse move. In VREX, an energy value that occurred at temperature T_j in the past is used, and following the attempted exchange, the occurrence of this energy value is removed from the potential energy list. This constitutes a "virtual exchange".

VREX is derived to obey detailed balance in a similar manner to RE.[2,14] Consider a state A in which a replica with configuration X is at temperature T_i and a virtual replica with configuration X_{virt} is at temperature T_j. An exchange is

attempted to a state B in which the replica with configuration X is at temperature T_j and the virtual replica is at temperature T_i. The detailed balance criterion is written

$$P(A) \, P(A \rightarrow B) = P(B) \, P(B \rightarrow A) \quad (11)$$

Transition probabilities that satisfy this criterion can then be derived as follows:

$$\frac{P(A \rightarrow B)}{P(B \rightarrow A)} = \frac{\dfrac{e^{-\beta_j H(X)}}{Z_j} \dfrac{e^{-\beta_i H(X_{\text{virt}})}}{Z_i}}{\dfrac{e^{-\beta_i H(X)}}{Z_i} \dfrac{e^{-\beta_j H(X_{\text{virt}})}}{Z_j}} = e^{-(\beta_j - \beta_i)(E(X) - E(X_{\text{virt}}))} \quad (12)$$

This transition probability is satisfied by the Metropolis criterion in eq 10. It also resembles the transition probability for RE (eq 8). The VREX algorithm is completely analogous to RE, except that one replica undergoes a virtual reverse move. Similarly, SREM also involves a virtual reverse move, in this case, by a replica whose energy distribution is represented by the PEDF. In contrast, the RE algorithm involves two real replicas undergoing temperature moves.

In practice, VREX requires very short initial simulations in order to generate a preliminary list of energies for each temperature. These lists are then updated as the simulation progresses, with values being added from each short MD simulation between exchange attempts and values being removed as they are used in virtual exchanges. It is possible to run out of potential energy values in the primary lists if temperatures are sampled heterogeneously. In order to address this possibility, implementations of VREX may include the use of secondary lists, to which potential energies from the primary lists are moved after a single use. Potential energies from a secondary list may be used in the rare case that the primary list for that temperature is completely used. Further, recent values can be prioritized in the primary lists, and relatively short and continually overwritten secondary lists can be maintained in order to reduce the likelihood of using pre-equilibration potential energies in post-equilibration virtual exchanges.

The main advantage of VREX is that it avoids the need to calculate converged PEDFs (like SREM) or weight factors (like ST) and only requires a short list of potential energies to begin sampling. It also addresses the main shortcoming of RE because it completely eliminates the synchronization between replicas, as well as the need for a fixed number of replicas. It is theoretically very similar to RE, with the addition of a variable time delay between the time when a potential energy is produced and when it is used for an exchange.

Distributed Replica Sampling (DR). Distributed replica sampling[11] is a general scheme for Boltzmann sampling of conformational space in which multiple replicas undergo a random walk in a reaction coordinate of interest. Individual replicas are coupled through a generalized Hamiltonian containing a potential energy term that depends on the distribution of all replicas, which acts to enforce a desired sampling distribution of the reaction coordinate. DR can therefore be used to enforce uniform sampling along a reaction coordinate of interest. This may be, for instance, a nonphysical spatial "fourth" dimension[12] or a dihedral angle.[13] We briefly summarize the DR algorithm with temperature as the coordinate.[11] The implementation of DR in other coordinates has also been previously outlined.[11−13]

The generalized Hamiltonian of DR in temperature contains a pseudoenergy term and depends on the current inverse temperature and current configuration (q) of all replicas:

$$H(q_1, \beta_1, q_2, \beta_2, ..., q_M, \beta_M) = \sum_{m=1}^{M} \beta_m \, \text{E}(q_m) + \\ \text{DRPE}(\beta_1, \beta_2, ..., \beta_M) \quad (13)$$

where E is the potential energy. There are M replicas in total, each labeled by an index $m = 1, ..., M$. The distributed replica potential energy (DRPE) can take any functional form that depends on the distribution of replicas and fulfills the purpose of enforcing homogeneous sampling of the temperature coordinate. Importantly, although the DRPE is a pseudoenergetic penalty, it is not a function of system complexity.[11] The probability of a replica currently at temperature T_i jumping to a temperature T_j is

$$p(T_i \rightarrow T_j) = \min \begin{cases} 1 \\ e^{-(\beta_j - \beta_i)E(q_i) - (\text{DRPE}_j - \text{DRPE}_i)} \end{cases} \quad (14)$$

which depends on the difference between the DRPE with the replica at temperature T_j (DRPE_j) and that at temperature T_i (DRPE_i).[11] DR can be analogously used to achieve a random walk in a parameter of the Hamiltonian, ξ, with an exchange probability:

$$p(\xi_i \rightarrow \xi_j) = \min \begin{cases} 1 \\ e^{-\beta[(H(q,\xi_j) - H(q,\xi_i)) + (\text{DRPE}_j - \text{DRPE}_i)]} \end{cases} \quad (15)$$

DR was designed specifically to suit shared or distributed computing platforms.[11] In contrast to RE, in which pairwise exchanges of replicas are attempted, DR considers stochastic moves of individual replicas one at a time. The stochastic move of one replica is coupled to the distribution of all other replicas through the DRPE, and no direct communication between replicas is required. In DR, synchronization of exchange attempts is therefore completely eliminated, which results in 100% CPU utilization.[11] The algorithm also readily accommodates fluctuations in CPU availability.[11] DR in combination with thermodynamic integration (TI) was shown to sample conformational space more effectively than TI alone in the calculation of the binding free energy of benzene to T4 lysozyme, while simultaneously optimizing the use of available computational resources.[12] This approach was also successfully employed to compute partial water occupancy in the pathway of proton uptake in cytochrome c oxidase.[50] In addition, DR has been combined with umbrella sampling (DRUS) to allow equilibrium exchange between different umbrella biasing potentials.[13,50] When applied to alanine dipeptide, umbrella sampling alone exhibited quasi-nonergodic behavior, while DRUS alleviated this systematic error.[13]

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2645**

Application of the DRPE restores sampling homogeneity of temperature only when the DRPE contribution is large enough to balance the preference for sampling the lowest temperature. When ST is conducted with all weight factors equal, as outlined in eq 14, all replicas migrate with a strong preference for the lowest temperature (as described above). In this case, a very strong DRPE is required to achieve sampling homogeneity. However, it has been demonstrated that, as the energetic penalty of the DRPE becomes stronger, replica mobility (as measured by acceptance ratio) decreases,[11] and therefore some modification to the DR exchange probability is necessary. This issue can be addressed by simply adding weight factors to the exchange probability, analogous to the weight factors, f, in the DRUS exchange probability:[13]

$$p(\xi_i \rightarrow \xi_j) = \min\left\{\begin{array}{l} 1 \\ e^{-\beta[(H(q,\xi_j)-H(q,\xi_i))-(f_j-f_i)+(\text{DRPE}_j-\text{DRPE}_i)]} \end{array}\right. \tag{16}$$

This form of the exchange probability results in good replica mobility and nearly perfect sampling homogeneity of the reaction coordinate.[13] The efficiency and practical advantages of DR in other coordinates have been well established.[11−13] It is therefore a central objective of this study to develop and test an implementation of DR which functions optimally in temperature space.

Simulated Tempering Distributed Replica Sampling (STDR). Building on the success of the both the ST method[3,4] and DR sampling,[11−13,50] we have developed a new algorithm, STDR, which combines the two approaches. STDR is essentially DR implemented rigorously in temperature. The combination of these two methods was originally suggested when DR was developed.[11] In STDR, approximately homogeneous sampling of a set of temperatures is enforced. The probability of accepting a move from a temperature T_i to a neighboring temperature T_j is

$$p(T_i \rightarrow T_j) = \min\left\{\begin{array}{l} 1 \\ e^{-(\beta_j-\beta_i)E+(a_j-a_i)-(\text{DRPE}_j-\text{DRPE}_i)} \end{array}\right. \tag{17}$$

This is the same as the exchange probability from ST, with the addition of the difference in DRPE between the states for which the replica is at temperature T_i (DRPE$_i$) and temperature T_j (DRPE$_j$). The calculation of the DRPE is straightforward. Its functional form depends upon the current temperatures of all replicas as follows:[12]

$$\text{DRPE} = c_1 \sum_{m=1}^{M} \sum_{n=1}^{M} [(\lambda_{m,\text{linear}} - \lambda_{n,\text{linear}}) - \omega(m-n)]^2 +$$
$$c_2 \left[\sum_{m=1}^{M} \lambda_{m,\text{linear}} - \omega \sum_{m=1}^{M} m\right]^2 \tag{18}$$

Replicas are labeled by indices m and n, where M is the number of replicas. The values of $\lambda_{m,\text{linear}}$ refer to a linearly spaced temperature coordinate. In this coordinate, the lowest temperature has $\lambda_{m,\text{linear}} = 1$, and the highest temperature has $\lambda_{m,\text{linear}}$ equal to the number of temperatures. This procedure transforms the exponentially spaced temperatures into a uniformly spaced coordinate. The factor ω, which we introduce to the DRPE in this work, is the ratio of the number of temperatures to the number of replicas. This factor allows DR to be used with an arbitrary number of replicas. The first term in eq 18 introduces an energetic penalty for two replicas sampling the same temperature, while the second term introduces a penalty for an overall drift of the replicas toward high or low temperatures. The second term is not essential when using DR in temperature. The constants c_1 and c_2 control the influence of the DRPE and can be tuned to enforce homogeneous temperature sampling as required.[11] In the case of accurate weight factors, the influence of the DRPE only needs to be small such that values of c_1 and c_2 near zero can be used. With increasingly inaccurate weight factors, larger DRPE values are required to maintain homogeneous temperature sampling, and this reduces the acceptance ratio to some degree. An example calculation of the DRPE using the temperature as the reaction coordinate is provided as Supporting Information.

If the weight factors, a_m, are inaccurate, ST results in uneven sampling of the temperature coordinate. As we will demonstrate, introducing the DRPE recovers homogeneous sampling. The STDR method is therefore more generally applicable than ST because it can make use of a poor estimate of the dimensionless Helmholtz free energies and still yield uniform sampling of the canonical ensembles at each temperature. STDR is equivalent to ST in the limit of one replica and is compatible with adaptive schemes for computing weight factors. Below, we show that STDR is the preferred method for systems with a complex energy landscape for which limitations on computational resources preclude obtaining sufficiently accurate estimates of Helmholtz free energies for a ST simulation.

Test System. For the purpose of comparing different generalized-ensemble methods, we use two related test systems, the peptides GVGVPGVG and (GVPGV)$_7$. These peptides are both based on the pentapeptide GVPGV, which is found as a repeat motif in the protein elastin.[51] In our previous study of (GVPGV)$_7$ and other related elastin-like peptides, we observed that this peptide is intrinsically disordered, having many conformations and no extended secondary structure in the form of α-helices or β-sheets.[52] Understanding the structural heterogeneity of elastin-like peptides is required to elucidate the structure−function relationship of elastin, for which experimental characterization is notoriously difficult due to its flexibility and insolubility. The peptide GVGVPGVG has also been studied previously and was suggested to exhibit an "inverse temperature transition" with an increased probability of "closed" conformations (in which the N and C termini are closer than 8 Å) at higher temperatures.[53] On the basis of this work, the octamer is a simple yet appropriate peptide to study in the aim of understanding the temperature-dependent behavior of elastin. Because the main focus of this paper is the thorough comparison of generalized-ensemble methods using these peptides as test systems, we do not elaborate fully on the structural details of either the octapeptide or the 35-residue peptide in this paper. A full characterization of the conformational landscape of these peptides will be the subject
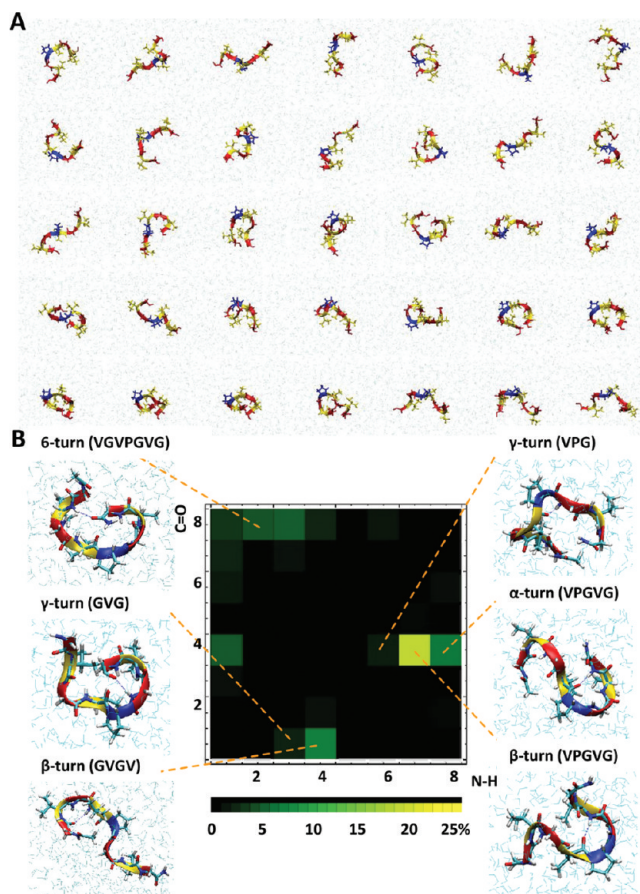
Figure 1. Conformational landscape and hydrogen-bonding contact map of GVGVPGVG. (A) A selection of 35 random conformations from the STDR simulation at 280 K, with glycine in red, valine in yellow, and proline in blue. (B) Hydrogen-bonding contact map at 280 K, with corresponding snapshots showing the presence of significantly populated contacts. N−H groups are on the horizontal axis, and C=O groups are on the vertical axis. Each square in the matrix (*i,j*) corresponds to a contact between the N−H group of residue *i* and the C=O group of residue *j*. The color scheme of the legend indicates the relationship between color and probability of contact formation.

of future work. Both GVGVPGVG and $(GVPGV)_7$ are valuable test systems because of their structural complexity and the fact that they represent a real scientific problem in the sense that they are not well understood or characterized a priori. Simple test systems are often used for comparison purposes, such as alanine dipeptide,[10,13] although generalized-ensemble methods are typically applied to systems which are much larger and more complex. While simple test systems are useful for the sake of demonstration and for the elucidation of major problems, they are less likely to detect the subtleties and practical issues experienced when studying systems of biologically relevant complexity.

The conformational landscape of the octapeptide is complex, with many energetically accessible states that must be sampled in order to accurately compute free energies. A representative selection of these conformations is shown in Figure 1A, with "closed" states in which the N and C termini are in close proximity, "hairpin"-like states, and extended structures. Although it is a short peptide, GVGVPGVG

represents a challenging sampling problem due to the large number of thermally accessible conformations. In Figure 1B, we show the hydrogen-bonding contact map for this peptide obtained using STDR. The only secondary structure consists of hydrogen-bonded turns, with no α-helix or β-sheet. The most populated turn is the VPGV β turn, with a hydrogen bond between the C=O group of valine 4 and the N−H group of valine 7. Several other turns form with lower populations. As we will show, single-temperature MD, if run for a sufficiently long time, provides a converged description of the conformational landscape. This makes it an ideal test system because we can verify that the generalized-ensemble algorithms, given sufficient sampling, lead to correct Boltzmann-weighted sampling of conformational space, in addition to an assessment of their relative efficiency.

The 35-residue peptide, $(GVPGV)_7$, is used as a more complex test system to demonstrate the sampling enhancement provided by STDR for a landscape which not only has many populated states but also has significant energetic barriers between those states. The larger system is only simulated using constant temperature MD, ST, and STDR because of the extensive amount of computational resources required. Of the methods we consider, STDR is better suited to this particular application on the basis of its performance for the octapeptide. It is as efficient and accurate as the other methods, while offering the most practical advantages for a large and complex system (see below).

Simulation Details. For all five methods (ST, STDR, SREM, RE, and VREX), the same exponentially spaced temperature list was used. This list is provided as Supporting Information Table S1. The simulation system consists of the GVGVPGVG octapeptide, capped with an acetyl group at the N terminus and an NH_2 group at the C terminus, in a $3 \times 3 \times 3$ nm^3 box with 872 water molecules. The same fully extended starting structure was used for all temperatures and all methods. Simulations were performed using the GROMACS MD simulation package, version 3.3.1,[54,55] with the OPLS-AA/L force field[56,57] for the solute and the TIP3P model for water.[58] Periodic boundary conditions were applied. The switch function of GROMACS was used for Lennard-Jones interactions, which corresponds to the usual Lennard-Jones function until 1.3 nm is reached, after which it is switched to reach zero at 1.4 nm. Covalent bonds involving hydrogen atoms were constrained with the SHAKE algorithm.[59] Calculations of electrostatic forces utilized the particle mesh Ewald summation method[60,61] with a Fourier spacing of 0.15 nm and a fourth-order interpolation. The real-space Coulombic cutoff was 1.49 nm. All MD simulations were performed in the canonical ensemble. Peptide and solvent were coupled to the same reference temperature bath with a time constant of 2 ps using the Nosé−Hoover method.[62,63] An integration step size of 2 fs was used, and coordinates were stored every 1 ps.

In order to compare the generalized-ensemble methods, the simulations were conducted as similarly as possible. To this end, the same total amount of simulation time (summed over all replicas) was performed. This amount was 4.75 μs, with an average of approximately 144 ns per replica. This time was used because it was sufficient

for all methods to achieve statistical convergence, as shown in the results. Stochastic exchanges using the Metropolis Monte Carlo algorithm[64] were attempted every 25 ps. Exchange probabilities were calculated using eqs 6, 8, 9, 10, and 17, as appropriate for the method. Details of the calculation of weight factors and PEDFs are discussed below. The constants c_1 and c_2 for the DRPE in eq 18 were both 0.005. These values were found to achieve an appropriate balance between homogeneity of temperature sampling and replica mobility.[11] The value of the factor ω was 1.0, since the number of replicas equaled the number of temperatures. The generalized-ensemble algorithms were implemented using an in-house bash script. Software for distributed replica sampling is also available online at www.pomeslab.com.

The same simulation protocol was used for the simulation of $(GVPGV)_7$, which was simulated in a $4.5 \times 4.5 \times 4.5$ nm^3 box with 2856 water molecules using both ST and STDR. Starting conformations and weight factors for each temperature were generated using canonical MD for 15 ns per temperature (storing 250 energy values per picosecond). A total of 70 temperatures were used for each generalized-ensemble simulation. The list of temperatures is provided in Supporting Information Table S1. Temperatures were spaced more closely than those of the octapeptide. This is because it is a larger system, resulting in less overlap between potential energy distributions of adjacent temperatures for a given temperature separation. This system was simulated for a total of 8.2 μs (117.6 ns per replica on average) using the STDR algorithm. An ST simulation using the same weight factors was also performed for 420 ns (6 ns per replica on average). ST and STDR simulations were also performed using weight factors calculated using the first 500 ps of continuous MD at each temperature for a total of 280 ns.

A conventional MD simulation of the 35-residue peptide system in the isothermal−isobaric ensemble was also performed using GROMACS, version 4.0.2.[65] In this simulation, a 4 fs time step was used, and constraints on bonds and angles involving hydrogen were imposed using the LINCS algorithm.[66] This simulation was run for 200 ns at 261 K, which corresponds to the lowest temperature in the STDR simulation. The pressure was kept constant at 1.0 bar using the Parrinello−Rahman algorithm.[67]

The analysis of the data accumulated in the trajectories was performed using an in-house script based on a modified version of the Dictionary of Secondary Structure in Proteins (DSSP).[68] For each snapshot, possible backbone hydrogen bonds were evaluated using both (a) the energetic criterion of DSSP and (b) the following geometric criteria: (i) donor−acceptor and hydrogen-acceptor distances are less than 3.5 and 2.5 Å, respectively, and (ii) the value of the acceptor−donor-hydrogen angle is less than 60°. Definitions of turns and bends are the same as those in DSSP.[68] End-to-end distance (EED) is calculated as the distance between the α carbons of the first and last residue. Root mean square deviation (rmsd) was calculated using the g_rms program in GROMACS.[65] All molecular visualizations in the manuscript were produced using VMD.[69]

Calculations of Weight Factors for ST and STDR, PEDFs for SREM, and Potential Energy Lists for VREX. The calculation of weight factors for eqs 6 and 17 required initial simulations of the octapeptide in the canonical ensemble for each of the temperatures listed in Supporting Information Table S1. These simulations were performed using conventional MD for 19.5 ns (for a total simulation time of 643.5 ns). Although obtaining these accurate weight factors was resource-intensive, it involved a straightforward procedure. The weight factors were computed using the average potential energy at each temperature according to eq 7.[30] The accuracy of these weight factors was assessed by using them in an ST simulation and observing the temperature sampling uniformity, as shown in the results. Since all temperatures were sampled with nearly equal probability, as expected from eqs 3 and 5 for accurate dimensionless Helmholtz free energies, these weight factors were deemed to be sufficiently converged and correct.

Using the same data from the conventional MD simulations, PEDFs were computed as described in the original SREM paper.[10] The convergence of the PEDFs was assessed by calculating the χ^2 measure suggested by Hagen *et al.*:[10]

$$\chi^2(t) = \sum_{n=1}^{N_{\text{bins}}} (P_i(t) - P_{i,\text{reference}})^2 \qquad (19)$$

This measure computes the deviation of each bin in the current distribution, $P_i(t)$, from a reference distribution, $P_{i,\text{reference}}$. The current distribution is cumulative, using the data up to time t. For the reference distributions, we used PEDFs computed using all of the data at each temperature. By this assessment, the PEDFs appeared to be stationary, as shown in Figure 2A. When χ^2 was plotted individually for each temperature, we also observed that each PEDF was stationary. However, an initial SREM simulation using these PEDFs resulted in nonuniform sampling of temperatures. We therefore proceeded to calculate the PEDFs using a different data set. We used the first 25 ns at each temperature of the RE simulation (for a total time of 825 ns), and these were the PEDFs used for the SREM simulation. While this procedure is similar to what would likely be done in practice with SREM, we emphasize that making this selection of PEDFs gave SREM somewhat of an advantage over ST, since more data were used in the initial simulation. The use of RE in the calculation of PEDFs is similar to REST.[36,38] Although REST results in faster convergence of the weight factors compared to conventional MD, it may be difficult or impossible to obtain access to the required number of homogeneous and dedicated CPUs for the initial RE simulation. Thus, we did not use REST to obtain the weight factors for ST to better represent the general case where it may not be convenient to do so. In contrast, it was necessary to use RE to obtain PEDFs for SREM in a reasonable amount of time.

Figure 2B shows the error in the exchange probability for both SREM and ST using the data from 19.5 ns of conventional MD at each temperature. The method for computing the error in exchange probabilities is provided in Appendix 1. The weight factors of ST produce an average

A

B

C

D
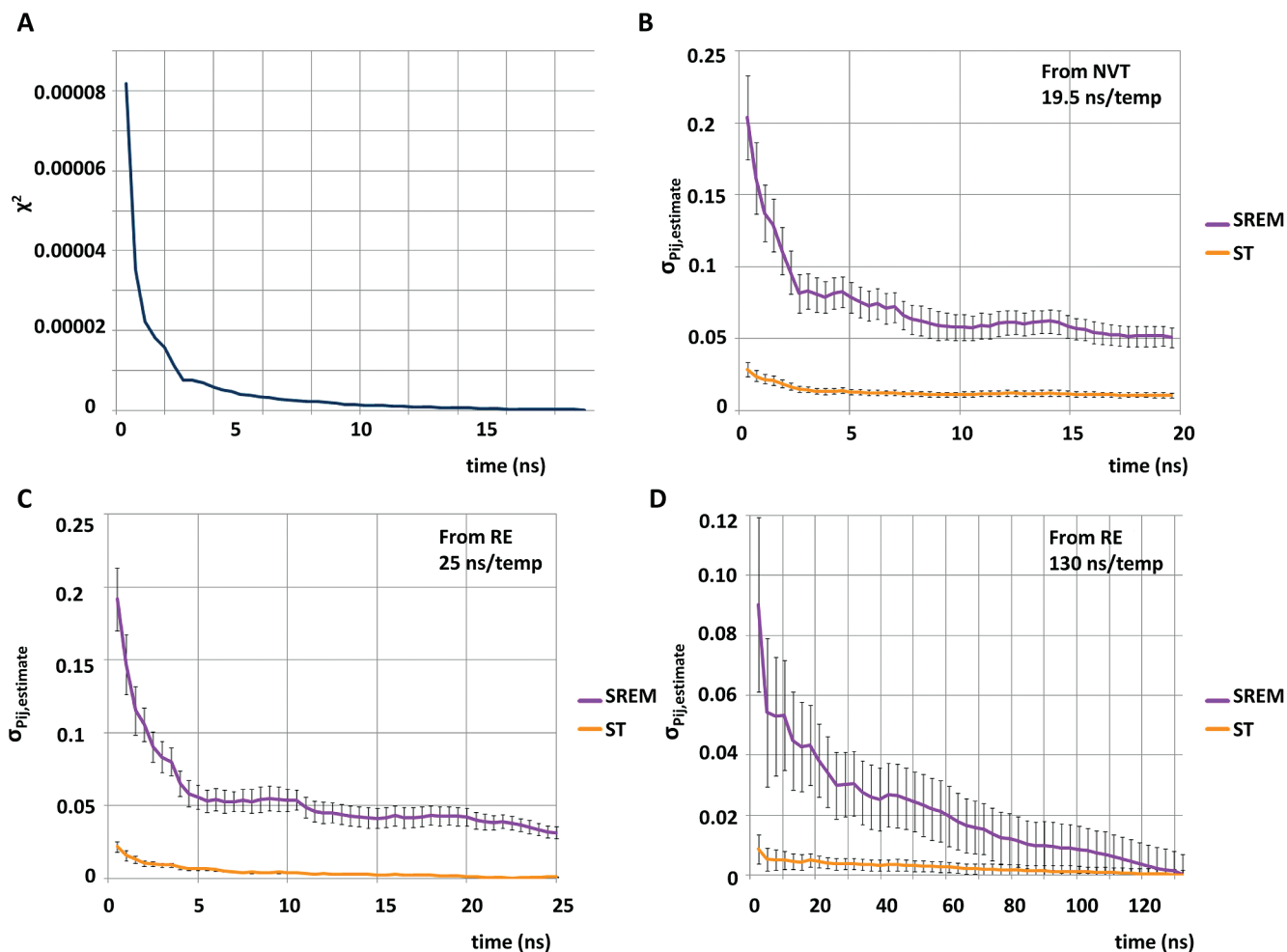
Figure 2. Assessing convergence of weight factors and PEDFs. (A) The convergence of the PEDFs for SREM is quantified using the χ^2 measure defined in eq 19. Using this measure, the PEDFs obtained using 19.5 ns of conventional MD at each temperature appear to be stationary. (B) Convergence of the PEDFs and weight factors using data from 19.5 ns of conventional MD per temperature, using the data from the complete RE simulation as a reference. (C) Similarly, data from the RE simulation, using only the first 25 ns per temperature, with the data from the complete RE simulation as a reference. (D) Similarly, using all of the data from the RE simulation (4.75 μs). Error in the acceptance ratio is shown in B, C, and D for both ST in yellow (computed using eqs 25 and 29) and SREM in purple (computed using eqs 24 and 27).

error in the exchange probability of less than 2% after 19.5 ns per temperature. Using the same amount of data, the PEDFs produce a significantly higher error in the exchange probability (more than 5%), which is why the weight factors used in ST from conventional MD produced more homogeneous sampling than the PEDFs. In Figure 2C, the error in the exchange probability for both ST and SREM is shown using the data from the first 25 ns at each temperature of RE. This data set was used to calculate the PEDFs for the SREM simulation, producing an error in the exchange probability of less than 4%. The convergence of the PEDFs estimated using all of the data from RE is shown in Figure 2D. The error in the exchange probability had only decreased to less than 2% after approximately 60 ns per temperature. That is, SREM would have required preliminary simulations which were half as computationally expensive as the entire RE simulation in order to produce error in the exchange probability equivalent to that of ST. The slow convergence of PEDFs is likely why they have been updated throughout

the course of the simulation in other studies.[10,14,47] However, an SREM simulation is strictly correct only with accurate PEDFs.[14]

Figure 2 demonstrates that the error in the weight factors of ST leads to smaller error in the exchange probability than the error in the average energy of PEDFs. This finding is in qualitative agreement with a previous study comparing SREM and ST for a helical peptide.[14] The PEDFs of SREM were observed to converge more slowly than the weight factors of ST when starting from a coil conformation, but not when both ST and SREM were started with a helical conformation.[14] In the original SREM paper, it was hypothesized, but not shown, that the calculation of PEDFs should be significantly easier than the calculation of weight factors for ST.[10] In fact, we observe that the opposite is true for this system. The weight factors converge significantly faster than PEDFs and lead to more homogeneous sampling of the temperature. The difference in errors is likely because the exchange probability in ST uses a difference in dimensionless free energies, whereas the absolute value of the potential

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2649**

energy is used in the exchange probability of SREM. Additionally, the method for computing the weight factors uses only the average potential energy at each temperature.[30] It is harder to converge a distribution of potential energies than the average of the distribution. Since weight factors are computed using average potential energies, they converge more quickly than PEDFs. The accuracy of the potential energy value selected from the discrete PEDF in SREM is also affected by the number of bins and the bin width.[10] The accuracy is decreased by having too few bins, whereas the convergence of the distribution is slower with a larger number of bins. These errors must therefore be balanced. Even if the PEDFs and weight factors converged at the same rate, ST has the advantage of convenience, since it entails storing a short list of weight factors rather than a distribution of energy values for each temperature.

We also tested the effects of using a poor estimate of the weight factors in ST. In order to generate suboptimal weight factors, we used the data from the first 750 ps of the RE simulation. This required a total of 24.75 ns summed over all temperatures, compared to 643.5 ns used to generate accurate weight factors. These weight factors produced inhomogeneous sampling of the temperature, confirming that they were inaccurate estimates of the dimensionless Helmholtz free energies (as is demonstrated in the results below). The purpose of this exercise was to emulate the more general case of a complex system for which one may not be able to accurately calculate weight factors due to the prohibitive computational cost. ST and STDR simulations carried out with these inaccurate weight factors will hereafter be referred to as STb and STDRb, respectively.

Potential energy lists for the VREX simulation were also generated using the RE data. A list of 1000 energy values from the first 1 ns was used for each temperature. We did not run out of potential energy values in the primary lists and therefore no secondary lists were used. However, secondary potential energy lists may be necessary in the application of this method to other systems. In summary, we highlight the varying costs of the initial simulations for each of the methods in terms of the simulations times: RE (0 ns), VREX (33 ns), SREM (825 ns), STDR (643.5 ns), STDRb (24.75 ns), ST (643.5 ns), and STb (24.75 ns).

Results and Discussion

Practical Implementation Issues. Before we begin a detailed comparison of the efficiency of the temperature-based generalized-ensemble methods, we briefly compare them with regard to the practical issues encountered in their implementation. A summary of this comparative discussion is provided in Table 1. Prior to beginning an enhanced sampling simulation, it is necessary to assess the available computational resources, including the number of processors available, the heterogeneity of their speeds, and their failure rate (frequency of "crashes").

In terms of the number of CPUs required, the RE algorithm specifies that the number of replicas equal the number of temperatures, which grows with system size. In the typical implementation of RE, the number of processors equals the

Table 1. Practical Advantages and Disadvantages of Generalized-Ensemble Algorithms[a]

Implementation Issue	RE	SREM	VREX	STDR	ST
Scalable to any number of CPU's (even one CPU)?					
Algorithm readily accommodates a fluctuating number of CPU's?					
Efficiency impaired by inhomogeneity of CPU's?					
Performance severely affected by CPU failure?					
Initial calculation of weight factors, potential energy distribution functions or potential energy lists required?					
Initial simulation times (ns)	0	825	33	643.5 (STDR) 24.75 (STDRb)	643.5 (ST) 24.75 (STb)

[a] If a method is not affected by an implementation issue, the corresponding square is colored in green. Yellow indicates that the issue is somewhat of a concern, and red indicates that it is potentially a major pitfall. The only major issues for SREM and ST are the calculation of PEDFs and weight factors, respectively. STDR and VREX are not severely affected by any implementation issue. However, they do require very short initial simulations to obtain weight factors and potential energy lists and, ideally, would not be run with only one replica. RE, in its typical implementation in which each replica is run on a dedicated CPU, is hindered by all of the issues listed, except that it does not require any initial simulation.

number of replicas. If it is not possible to obtain access to the required number of processors, an alternative method or a more advanced RE implementation must be sought. Another possible scenario is that extra processors are available, which could be utilized to speed up the calculation, but the RE algorithm does not allow the possibility of having more replicas than temperatures. This particular issue is addressed by the MREM algorithm, which utilizes multiple layers of replicas.[23] However, there is no general mechanism to adapt RE to use available resources most efficiently. In contrast, both ST and SREM completely eliminate the need for a specific number of replicas. Multiple ST or SREM simulations can be run independently to take advantage of a computing cluster or distributed computing. The benefit of utilizing several processors simultaneously, each running an independent ST or SREM simulation, is simply reaching convergence more quickly in terms of wall clock time. Similarly, STDR and VREX algorithms do not require a fixed number of replicas. However, the aim of the DRPE is to enforce homogeneous sampling of temperatures for multiple replicas. Using only one replica is therefore not optimal, and ideally one would use a number of replicas comparable to the number of temperatures, though there is no specific requirement. A VREX simulation can in principle have any number of replicas. However, there is likely some benefit to having multiple replicas sampling different regions of conformational space in the updating of the potential energy lists (that is, running more than one replica at a time).

Of the generalized-ensemble methods we consider, only RE prevents the number of replicas from fluctuating during the course of the simulation. This may be a drawback in distributed

computing platforms and shared computing clusters, for which there is no way to predict the number of available processors in advance. Furthermore, the efficiency of RE is significantly affected by inhomogeneity of CPU speeds. Each exchange step can only occur when all of the replicas have completed their MD calculation. Any inhomogeneity in the computing environment results in a waste of computational resources, as some replicas must wait for the replica with the slowest processor to finish its calculation. This issue has been partly addressed by the asynchronous replica exchange method, although some degree of synchronization is still required for the replicas undergoing exchange.[43] Since none of the other methods require any direct communication between replicas, they do not suffer from this inefficiency. Another key drawback of the typical implementation of RE is its sensitivity to CPU failure.[10,11] If one of the replicas is running on a processor that crashes, the entire RE simulation is stalled until this replica can be restarted on a functioning processor. The time wasted due to CPU failure depends on the failure rate of the cluster and can be quite significant. Failure rates also rise with the number of replicas, and therefore the failure rate of RE is equal to the number of replicas times the failure rate of either SREM or ST.[10]

In contrast to RE, the other four methods all have the advantage of not requiring a fixed and synchronized cluster of CPUs to function optimally. From a practical point of view, these methods are all superior to RE, except in one regard. Only RE does not require initial simulations at multiple temperatures to obtain weight factors, PEDFs, or potential energy lists. In particular, ST and SREM appear to only be suited to systems for which accurate weight factors or PEDFs can be calculated in a reasonable amount of simulation time. For the test system in the present study, accurate weight factors for ST were computed using single-temperature MD, whereas SREM required more simulation time and the use of RE in order to obtain sufficiently accurate PEDFs (see methods section). We will demonstrate that STDR can function with less accurate weight factors and therefore requires less initial simulation time than ST. Finally, VREX requires significantly less initial simulation than SREM, STDR, or ST. Only short lists of potential energies at each temperature are needed to begin a VREX simulation.

An ideal temperature-based generalized-ensemble method would not require a significant initial simulation (as do ST and SREM) but also would not involve the use of a large cluster of homogeneous CPUs (as is common for RE). STDR and VREX address both of these issues, and are the most flexible algorithms in terms of practical concerns. These issues are particularly important if one is using a distributed computing platform with fluctuating numbers of heterogeneous CPUs in many different locations, or a shared computing cluster, which may present similar limitations.

Diffusion in Temperature. We characterize the efficiency of the temperature diffusion of each method using several different metrics, which are summarized in Table 2. First, we calculate the average acceptance ratio, which is a metric commonly reported for RE simulations.[70] The methods separate into two categories based on their acceptance ratios: the RE-based methods (RE, VREX, and SREM) and the ST-based methods (ST, STb, STDR, and STDRb). ST has a

Table 2. Evaluating Temperature Diffusion

Property	RE	VREX	SREM	STDR	STDRb	ST	STb
Acceptance Ratio	0.237	0.241	0.248	0.378	0.376	0.463	0.404
	7	6	5	3	4	1	2
Replica Speed (distance/time)	0.058	0.050	0.051	0.058	0.059	0.065	0.065
	4	7	6	4	3	1	1
Mean Free Path	0.322	0.245	0.255	0.402	0.402	0.431	0.445
	5	7	6	3	3	2	1
Diffusion Coefficient	0.208	0.181	0.156	0.195	0.196	0.246	0.249
	3	6	7	5	4	2	1
Average Deviation from Sampling Homogeneity (%)	0	6.62	12.61	2.50	2.98	3.81	17.40
	1	5	6	2	3	4	7
Composite Score (a normalized linear combination)	0.79	0.70	0.68	0.87	0.88	0.98	0.94
	5	6	7	4	3	1	2

The quality of the random walk in temperature for each generalized-ensemble method is assessed using five criteria, and an overall score is obtained by taking the normalized linear combination. Normalization is performed by dividing each value by the maximum value of that measure. The fifth measure was included in the normalized linear combination as $1.0 -$ (average deviation from sampling homogeneity)/100% for consistency with the other measures of temperature diffusion. The overall scores for each property are ranked from 1 to 7, representing decreasing performance. Scores are colored as follows: 1 and 2 (green), $3-5$ (yellow), and 6 and 7 (red).

higher acceptance ratio than RE for the same set of temperatures, in agreement with a previous comparison of the methods.[36] Similarly, ST has a higher acceptance ratio than SREM.[14] Zhang and Ma also observed that the rate of traversing temperatures is faster in ST, and that this effect becomes especially apparent if separations between adjacent temperatures are large, or if exchanges are attempted less frequently.[32] Park proved that this is generally true for a given set of temperatures[37] and concluded with a question as to whether the enhanced acceptance ratio affects the rate of sampling different microstates, and therefore structural convergence. We investigate whether the higher acceptance ratios in serial tempering algorithms (both ST and STDR) compared to those in parallel tempering (RE, VREX, and SREM) do in fact lead to faster structural convergence in the next section. It should be noted that the DRPE in STDR decreases the acceptance ratio relative to ST, since it increases the probability of rejecting moves that result in inhomogeneous temperature sampling. The extent of this effect depends on the constants c_1 and c_2 in eq 18.[11]

Next, we consider a quantity which we call "replica speed". Back exchanges can occur in which a replica accepts a move to an adjacent temperature and at the next exchange returns to its previous position. These back exchanges contribute to the acceptance ratio, but they result in no net change in temperature, and typically no crossing of significant energetic barriers. In order to account for these "unproductive" moves, we calculate the replica speed as the average distance traveled after 50 exchange attempts. Values of replica speed are reported in Table 2. All of the methods have similar values for the replica speed, with SREM and VREX being slightly slower. The higher acceptance ratios of the ST-based methods do not correspond to significantly

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2651**

faster replica speeds. That is to say, the higher acceptance ratios for the ST-based methods are partly due to an increased frequency of unproductive moves.

Making an analogy with the replicas traveling in temperature space as a type of diffusion in a one-dimensional coordinate, we calculate the mean free path and diffusion coefficient for each method. Mean free path is defined as the average distance traveled between successive rejected moves ("collisions"). The diffusion coefficient is defined as the rate of change of the mean squared deviation of distance over time. We notice that ST, both with accurate and inaccurate weight factors, has the highest mean free path and diffusion coefficient. Both STDR simulations behave remarkably similarly and are slightly slower at diffusion in temperature compared to ST. RE has a higher diffusion coefficient than STDR, but a lower mean free path. It is also slightly more efficient at temperature diffusion than VREX or SREM.

Another important criterion is the deviation from sampling homogeneity, which indicates the amount of deviation from uniform sampling averaged over all of the temperatures:

$$\text{average deviation from homogeneity} = \frac{1}{M} \sum_{m=1}^{M} \frac{|N_m - \langle N_m \rangle|}{\langle N_m \rangle} \quad (20)$$

where the number of samples at temperature m is N_m, the average number of samples per temperature is $\langle N_m \rangle$, and M is the number of temperatures. We report the deviation from sampling homogeneity for each temperature in Supporting Information Figure S1. The coupling of upward and downward moves in the RE algorithm results in perfectly uniform sampling of all temperatures. STDR produces nearly uniform sampling, with deviations from uniformity of 2.50% and 2.98% for accurate (STDR) and inaccurate (STDRb) weight factors, respectively. This is expected because the application of the DRPE favors uniform sampling of the temperature coordinate.[11] Even with inaccurate weight factors, the temperature sampling is still uniform, and the diffusion coefficient is still approximately the same. This indicates that STDR in the general case (i.e., with inaccurate weight factors) still successfully produces uniform sampling and good mobility in temperature. Our ST simulation also results in nearly uniform sampling, confirming the accuracy of the weight factors. ST with inaccurate weight factors (STb) produces the least uniform sampling, by design (17.40%). We intentionally selected weight factors to produce uneven sampling to represent a more complex system for which calculating weight factors accurately would be computationally expensive. VREX produces relatively uniform sampling, with an average deviation of 6.62%. Most of the inhomogeneous sampling in VREX occurs early in the simulation when the potential energy lists were based on a small amount of sampling, and the sampling is increasingly homogeneous with time. In contrast, SREM does not produce uniform sampling, with less sampling at the lowest temperatures and an average deviation of 12.61%.

Temperature sampling efficiency is characterized by an overall score. The five measures of efficiency defined in this
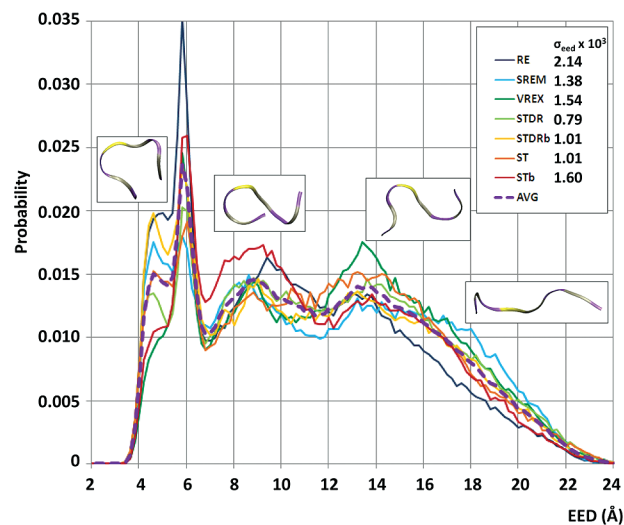
Figure 3. Assessing the accuracy of the EED distribution. The EED probability distribution is shown for each method with colors indicated in the legend using data from 280 K. The average distribution is computed as the average of all seven methods and is shown in purple (dashed line). The error of the distribution of each generalized-ensemble algorithm, σ_{eed}, is shown next to the legend and was computed using eq 21 with the average distribution as the reference.

section are combined by averaging their normalized values. The overall score for each method is reported in the last row of Table 2. ST with accurate weight factors performs the best overall, and all of the ST-based methods perform better than the RE-based methods (especially SREM, which has the lowest overall temperature diffusion score).

Convergence of Structural Properties. The octamer GVGVPGVG is a disordered peptide with many thermally accessible conformations, as shown in Figure 1. A useful descriptor of the conformation of such a short and flexible peptide is the end-to-end distance, EED. The EED probability distributions obtained using each of the generalized-ensemble methods at 280 K (the lowest temperature) are shown in Figure 3. Also shown is the average distribution, which is obtained by taking the average of all of the methods. There is no systematic bias of either ST-based or RE-based methods toward sampling particular conformations. For example, for the peak at 5 Å, RE, STDRb, and SREM are above the average while STDR, VREX, and STb are below. On the basis of this observation and of the large amount of sampling in the combined data set of all seven methods (nearly 35 μs), we take the average to be the "gold standard" for comparison throughout the analysis of structural convergence (it is hereafter referred to as the "reference"). We quantify the deviation σ_{eed}, of the EED distribution of each method $P_{\text{eed}}(n)$, from the reference EED distribution $P_{\text{eed,reference}}(n)$, by computing

$$\sigma_{\text{eed}} = \sum_{n=1}^{N_{\text{bins}}} (P_{\text{eed}}(n) - P_{\text{eed,reference}}(n))^2 \quad (21)$$

where the index n labels bins, and there are N_{bins} in total. The values of σ_{eed} are reported in Figure 3. STDR exhibits the best agreement with the average distribution. In
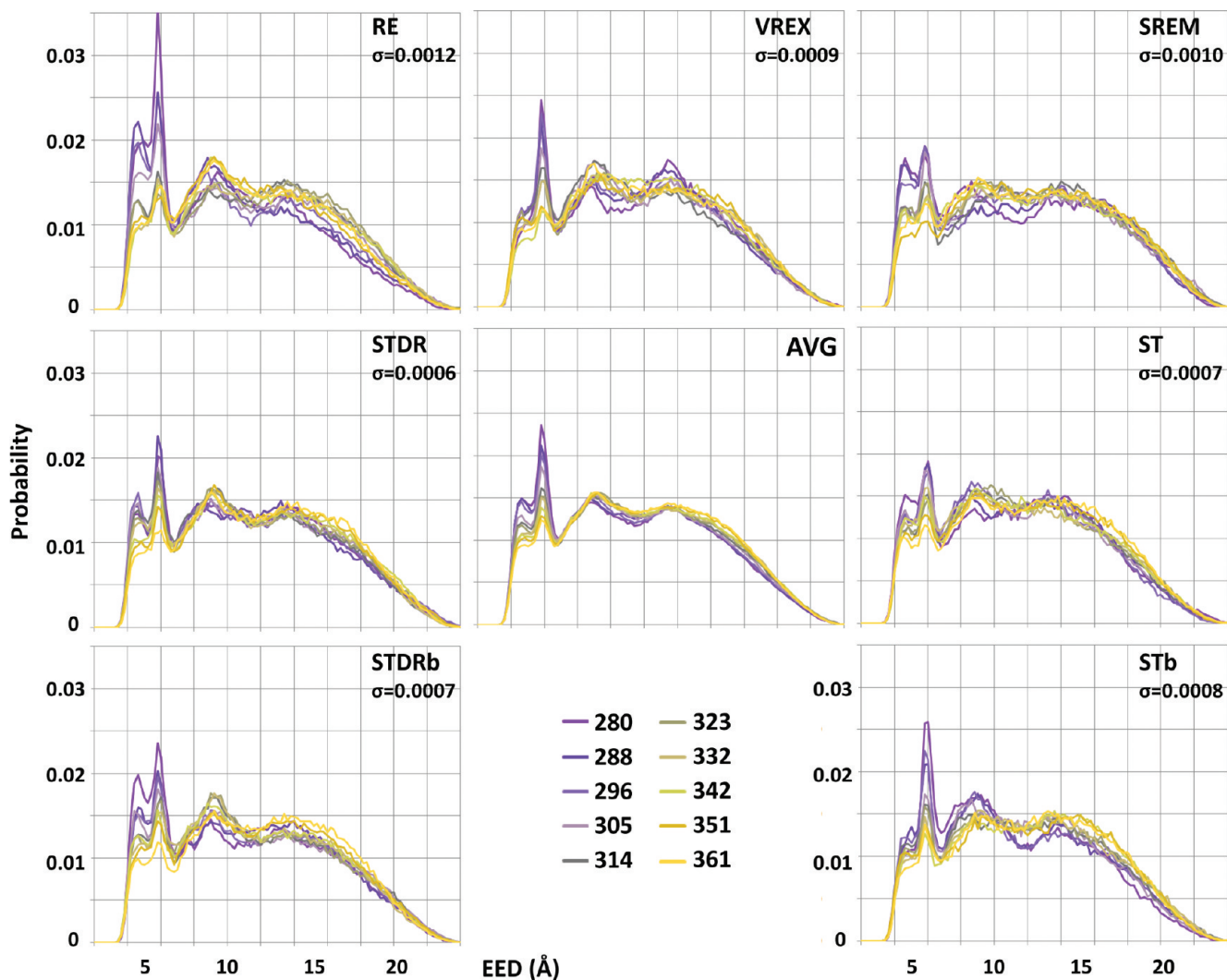
Figure 4. EED distributions at different temperatures. The EED probability distributions are shown for the lowest 10 temperatures for each generalized-ensemble method, as well as the average of all seven methods. The average error of the distributions of each generalized-ensemble algorithm, σ, is also shown. This was computed for each of the 10 temperatures using eq 21 with the average distribution as the reference, and the average of these errors is shown on each plot. The RE-based methods are shown in the top row and have larger errors than the ST-based methods, shown in the second and third rows. The average over all methods is shown in the central plot.

general, the ST-based methods have lower values for σ_{eed}, corresponding to more accurate EED distributions than the RE-based methods.

In order to confirm that the ST-based methods produce more accurate EED distributions when compared to the RE-based methods, EED distributions for the lowest 10 temperatures for each generalized-ensemble method are also computed and compared to the reference using eq 21. The EED distributions for each method and each temperature are displayed in Figure 4, along with the σ_{eed} value which is the average of the 10 temperatures. The ST-based methods produce EED distributions which are quantitatively more accurate than the RE-based methods at all temperatures. STDR shows the best overall agreement with the reference data set, with an average σ_{eed} value of only 0.006, and distributions which clearly show the same temperature trend as the reference distributions.

For a systematic comparison of the generalized-ensemble methods, the convergence of several structural properties in

addition to the EED distribution is considered. A useful ergodic measure is the 1,4 pair distance metric,[71,72] which quantifies the convergence of the distance between 1,4 residue pairs (residues with indices i and $i + 3$) over time. We extend this metric to include all residue pairs and therefore quantify the convergence of the α-carbon distance matrix as follows:

$$d_{dC\alpha matrix}(t) = \frac{1}{(N_{residues} - 1)!} \sum_{i=1}^{N_{residues}} \sum_{j=1}^{N_{residues}} (\bar{r}_{ij}(t) - \bar{r}_{ij,reference})^2$$

(22)

where the indices i and j correspond to residue number and the number of residues is $N_{residues}$. The difference between each average pairwise distance $\bar{r}_{ij}$ and the same average pairwise distance from the reference α-carbon distance matrix ($\bar{r}_{ij,reference}$) is computed. In this equation, t refers to simulation time accumulated at the temperature considered, and $\bar{r}_{ij}(t)$ is a cumulative average. As with EED, the average of all seven
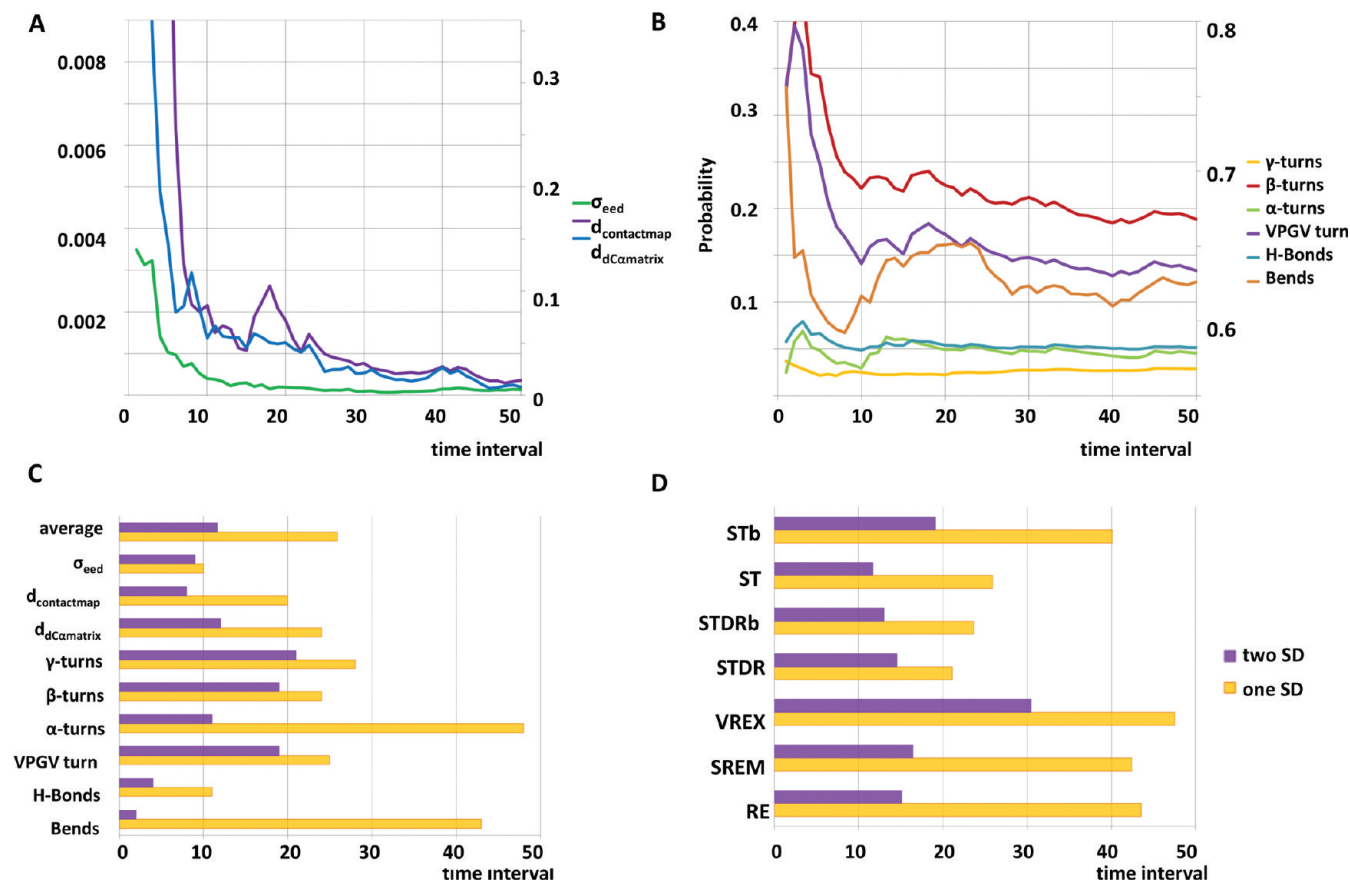
Figure 5. Assessing structural convergence using multiple criteria. The data for A, B, and C are from ST at 280 K. The trajectory is separated into 50 time intervals, and the quantities reported are calculated cumulatively. Time intervals are used to compare all methods fairly, since each method results in a different amount of sampling time at the lowest temperature. (A) Structural convergence is assessed using σ_{eed} (eq 21), $d_{contactmap}$ (eq 23), and $d_{dC\alpha matrix}$ (eq 22, plotted on the secondary axis). (B) The probability per residue of a γ-turn, β-turn, and α-turn are shown, as well as the population of the VPGV β-turn. The probability of a hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown. (C) For each of the structural properties shown in A and B, the time intervals at which they reached and remained within one and two standard deviations of the reference data set are shown. The average of these times is also shown, corresponding to the average structural convergence time, $\langle t_{sc} \rangle$. One standard deviation is calculated on the basis of the values of each of the seven generalized-ensemble methods at the end of the simulation and their standard deviation from the reference value. (D) The average structural convergence times for one and two standard deviations are shown for all seven methods at 280 K in yellow and purple, respectively. These times are provided for temperatures 288, 296, 305, 314, 323, and 332 K in Supporting Information Figure S2.

generalized-ensemble methods is used as the reference. We compute an analogous measure of convergence for the hydrogen-bonding contact map, depicted in Figure 1b:

$$d_{contactmap}(t) = \frac{1}{N_{residues}^2} \sum_{i=1}^{N_{residues}} \sum_{j=1}^{N_{residues}} (P_{ij}(t) - P_{ij,reference})^2$$

(23)

where P_{ij} is the probability of a hydrogen bond forming between the C=O group of residue i and the N−H group of residue j, and $P_{ij}(t)$ is a cumulative average of all of the data. The elements of the reference contact map, $P_{ij,reference}$, are computed using the data from all seven methods. We also directly compute the probability of forming certain turns (γ, β, and α turns, defined by hydrogen bonds between residues i and $i + 2$, $i + 3$, and $i + 4$, respectively) as well as the VPGV β-turn (shown in Figure 1, the most probable turn). In addition, the average probabilities of forming a hydrogen

bond and a "bend" (as defined in the DSSP algorithm[68]) on a per-residue basis are computed. The convergence of each of these structural properties is considered individually and compared to the reference data. Taken as a set, these structural properties provide a detailed description of the octapeptide's complex conformational ensemble.

A representative example of how these structural properties measure convergence is shown in Figure 5 for ST at 280 K. The convergence of the α-carbon distance matrix, the hydrogen-bonding contact map, and the EED distribution are displayed in Figure 5A. The cumulative averages for the different types of turns, as well as hydrogen bonds and bends, are shown in Figure 5B. It is apparent from both of these plots that selecting a particular time at which the simulation has converged is ambiguous. Each structural property appears converged at a slightly different time. This ambiguity highlights the importance of considering multiple metrics when discussing the convergence of a simulation. In order

to define convergence quantitatively, we consider the time taken to reach the reference value of the structural property of interest and remain within both one and two standard deviations, shown in Figure 5C. Taking the average of these times provides a composite measure of when structural convergence is reached, and this average is a "structural convergence time", t_{sc}. By comparing to the reference data, both convergence and accuracy are simultaneously assessed. The time at which each structural metric reaches the reference value is significantly different. For example, the EED distribution reaches the reference distribution faster than any of the other structural metrics, while the population of α-turns requires nearly the entire ST simulation to reach the reference value to within one standard deviation. The structural convergence times are provided in Figure 5D for each of the generalized-ensemble methods at 280 K. At this temperature, STDR converges fastest to the reference data, closely followed by ST and STDRb.

For a systematic ranking of the structural convergence times, t_{sc} is also calculated for the lowest seven temperatures. These times are provided in Supporting Information Figure S2. While STDR converges faster than the other methods at 280 K, this is not a general trend for all temperatures. Each temperature has a different t_{sc} for each method. The ranking of the methods varies between temperatures. For example, at 305 K, RE reaches convergence faster than STDR. This highlights the importance of evaluating more than the lowest temperature when comparing the performance of the methods, in addition to considering several structural metrics. It also suggests a way of quantifying the error in the measurement of t_{sc}. An average structural convergence time, $\langle t_{sc} \rangle$, for each method is obtained by averaging t_{sc} for the lowest seven temperatures, for both one and two standard deviations. The error in $\langle t_{sc} \rangle$ is then the standard error of these measurements. Figure 6A shows a two-dimensional plot of $\langle t_{sc} \rangle$ to within two standard deviations versus $\langle t_{sc} \rangle$ to within one standard deviation. Lower values for $\langle t_{sc} \rangle$ indicate faster structural convergence. A clear trend emerges: ST-based methods reach structural convergence more quickly than RE-based methods. The method that reaches convergence the fastest is ST with accurate weight factors, while the method slowest to converge is SREM. It is not possible to conclusively rank the other methods due to error in $\langle t_{sc} \rangle$. However, it is important to note that both VREX and RE converge faster than SREM. VREX is therefore not only a more convenient method for removing the synchronization from the RE algorithm than SREM, but is also faster at conformational sampling.

We can now answer a key question: does faster diffusion in temperature lead to a corresponding speedup in conformational sampling? Figure 6B demonstrates that this is in fact the case. The combined average structural convergence time, obtained by taking the sum of $\langle t_{sc} \rangle$ for one and two standard deviations, is plotted versus the composite temperature diffusion score from Table 2. The ST-based methods, which have higher acceptance ratios and diffusion coefficients, also exhibit faster structural convergence. This key observation indicates that, in general, it is preferable to use a ST-based method because it provides enhanced efficiency
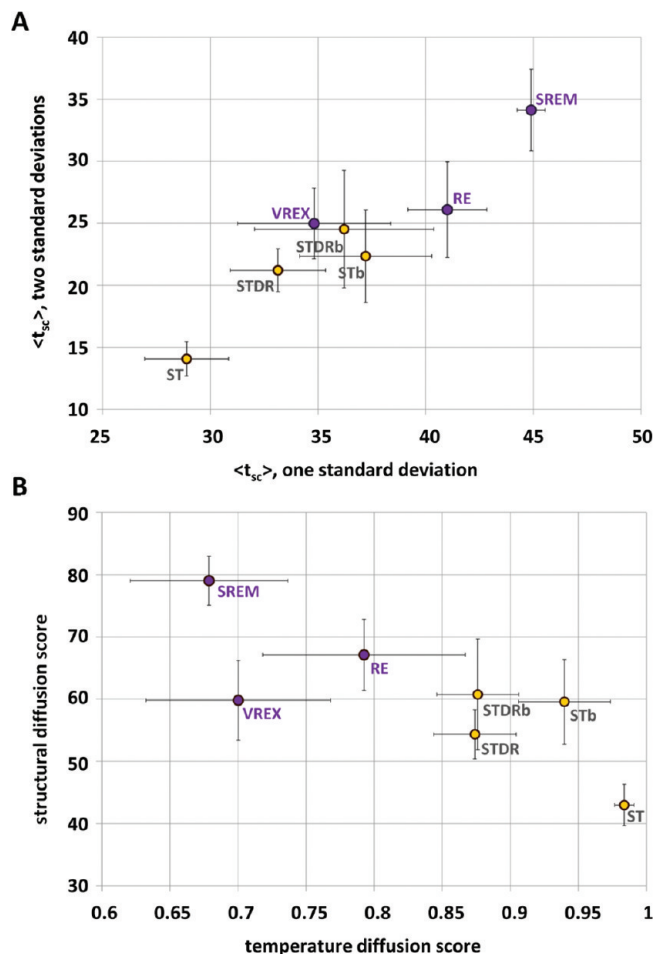
Figure 6. Correlation between structural convergence and temperature diffusion. (A) Average structural convergence times, $\langle t_{sc} \rangle$, obtained using the lowest seven temperatures are shown. The $\langle t_{sc} \rangle$ to reach two standard deviations is plotted against the $\langle t_{sc} \rangle$ to reach one standard deviation for each method. Error bars represent the standard error of $\langle t_{sc} \rangle$ for the seven temperatures. Another version of this plot is provided as Supporting Information Figure S3, with the $\langle t_{sc} \rangle$ for each temperature shown. (B) The $\langle t_{sc} \rangle$ times for one and two standard deviations from A are added together to create a structural convergence score, which is plotted against the temperature diffusion score from Table 2 for each method. A correlation is observed between structural convergence and temperature diffusion. ST-based methods (in yellow) have superior temperature diffusion, which leads to faster structural convergence compared to RE-based methods (in purple).

in terms of conformational sampling. ST with accurate weight factors is clearly superior in both temperature diffusion and structural convergence, while SREM is the least efficient method in terms of both of these metrics. In the case of a simple system for which weight factors can be obtained accurately with relatively little computational expense, ST is the method of choice. In the case of a more complex system for which sufficiently accurate weight factors might be expensive to obtain, the best choice would be to compute an initial estimate for the weight factors and use ST or STDR (corresponding to STb and STDRb here). Using the octapeptide as a test system, it is not possible to conclude which of these options is preferable. Inaccurate weight factors for this

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2655**

system yield comparable temperature diffusion and structural convergence for both STb and STDRb. To investigate this issue further, a more complex system, (GVPGV)$_7$, is also studied below.

Finally, another important question is whether inaccurate weight factors or PEDFs still lead to accurate, Boltzmann-weighted sampling at each temperature. It has been suggested that simulations with incorrect weight factors still yield correct statistics, only with suboptimal sampling of the temperature.[4] Analysis of the effect of suboptimal Helmholtz free energies on the accuracy of the data demonstrates that the resulting conformational populations are not biased by the use of inaccurate free energies (Figures 3 and 6a). Both ST and STDR with inaccurate weight factors (STb and STDRb) converge to the reference data set, which indicates that they achieve accurate, Boltzmann-weighted conformational sampling. Figure 6A shows that, even with inaccurate PEDFs, SREM still leads to Boltzmann-weighted sampling of conformational space (within one standard deviation) for this system. However, it converges more slowly than RE and all of the other generalized-ensemble algorithms considered in this study.

It has been pointed out that SREM is not rigorously correct when employing PEDFs that are not representative of the potential energies sampled during the simulation.[10,14] In other words, SREM is correct only when stationary potential energy distributions are used, which may in the general case require adaptation of the PEDFs. In contrast, VREX is inherently adaptive due to the update of the potential energy lists. All RE-based methods, including VREX and SREM, require an initial equilibration phase to reach stationary potential energy distributions. Quasi-nonergodicity due to finite run length is a potential issue for any of the generalized-ensemble algorithms but is most significant for SREM. Systematic error associated with selecting potential energies from unrepresentative conformations can be minimized in VREX by maintaining short, frequently updated potential energy lists. Reducing the size of the potential energy list in the VREX approach decreases the equilibration lag. Both VREX and RE use recent potential energy values for attempted moves. In addition, another inherent advantage of VREX over SREM is that, in VREX, potential energy values utilized in the virtual temperature transitions are drawn from the same ensemble as those utilized in the evaluation of actual transitions, albeit with a stochastic time delay.

Comparison of STDR and Conventional MD. The relative sampling enhancement of RE compared to conventional MD has been the subject of significant controversy.[20] For example, one study found that RE produced a speedup of 71.5 times at 275 K for a 21-residue helical peptide with implicit solvent, based on the autocorrelation function of helicity.[19] In another work, an RE simulation of met-enkephalin in explicit solvent sampled 5 times more conformational space than a conventional MD simulation of the same duration.[18] It has also been shown analytically that the expected speedup of RE is directly related to the activation enthalpy for two-state protein folding. The efficiency of RE is optimal when the maximum temperature is chosen just slightly above the temperature at which the folding activation

enthalpy is zero.[20] There are several key issues that emerge when reviewing comparative studies of RE and conventional MD. First, the observed sampling enhancement, or lack thereof, is heavily system-dependent, as well as dependent on the structural or thermodynamic parameter on which the comparison is based. Second, an evaluation of convergence for either the RE simulation or the MD simulation is often neglected. Comparisons of other generalized-ensemble methods, including ST and SREM, and conventional MD have also been performed.[14]

Here, we attempt to provide a rigorous comparison between STDR and conventional MD for the octapeptide (Figure 7). Figure 7A and B show a superposition of 200 structures obtained using STDR and MD, respectively, at 280 K. The amount of simulation time is the same for both methods (144 ns for conventional MD and 144 ns in total for all temperatures for STDR, corresponding to 4.4 ns at 280 K). The root-mean-square deviations (rmsd) of these two collections of structures, 3.52 Å for STDR and 3.88 Å for conventional MD, are comparable. By this measure, both STDR and conventional MD produce a similar amount of conformational sampling using the same amount of CPU time.

We also show the convergence of the structural properties described in the previous section for both STDR (in Figure 8C and E) and conventional MD (in Figure 7D and F). STDR converges more quickly, approximately by a factor of 2−3 at 280 K. However, given that STDR requires sampling 33 temperatures for the same amount of time, it is much less computationally efficient. Specifically, the STDR simulation was 4.75 μs, compared to 200 ns for conventional MD. Overall, for this particular system, there is no computational advantage in using STDR over conventional MD when the total cost of simulating all temperatures is considered.

However, in the present case, we are interested in the conformational ensemble at both low and high temperatures because of the predicted temperature transition of the octapeptide GVGVPGVG.[53,73,74] It is therefore still beneficial to use STDR because it enhances sampling at the individual temperatures. It is of key importance to note that we only know that conventional MD is able to satisfactorily reproduce the conformational ensemble of the octamer by simultaneously using generalized-ensemble algorithms. It is only by comparing to STDR, as well as the combined data set of all the generalized-ensemble methods, that we are able to verify the convergence of the conventional MD simulation. Pseudoconvergence can be observed for a structural ensemble generated by conventional MD which is energetically trapped.[13] In this way, it is possible to achieve convergence without simultaneously achieving accuracy. Using a generalized-ensemble method and allowing a random walk in temperature allows the system to overcome energetic barriers. Without knowledge of the energy landscape of the system of interest, it is hard to predict the expected sampling enhancement of a generalized-ensemble method. Similarly, it is hard to assess the accuracy of an apparently converged value, which is also expected to depend on the topology of the energy landscape.
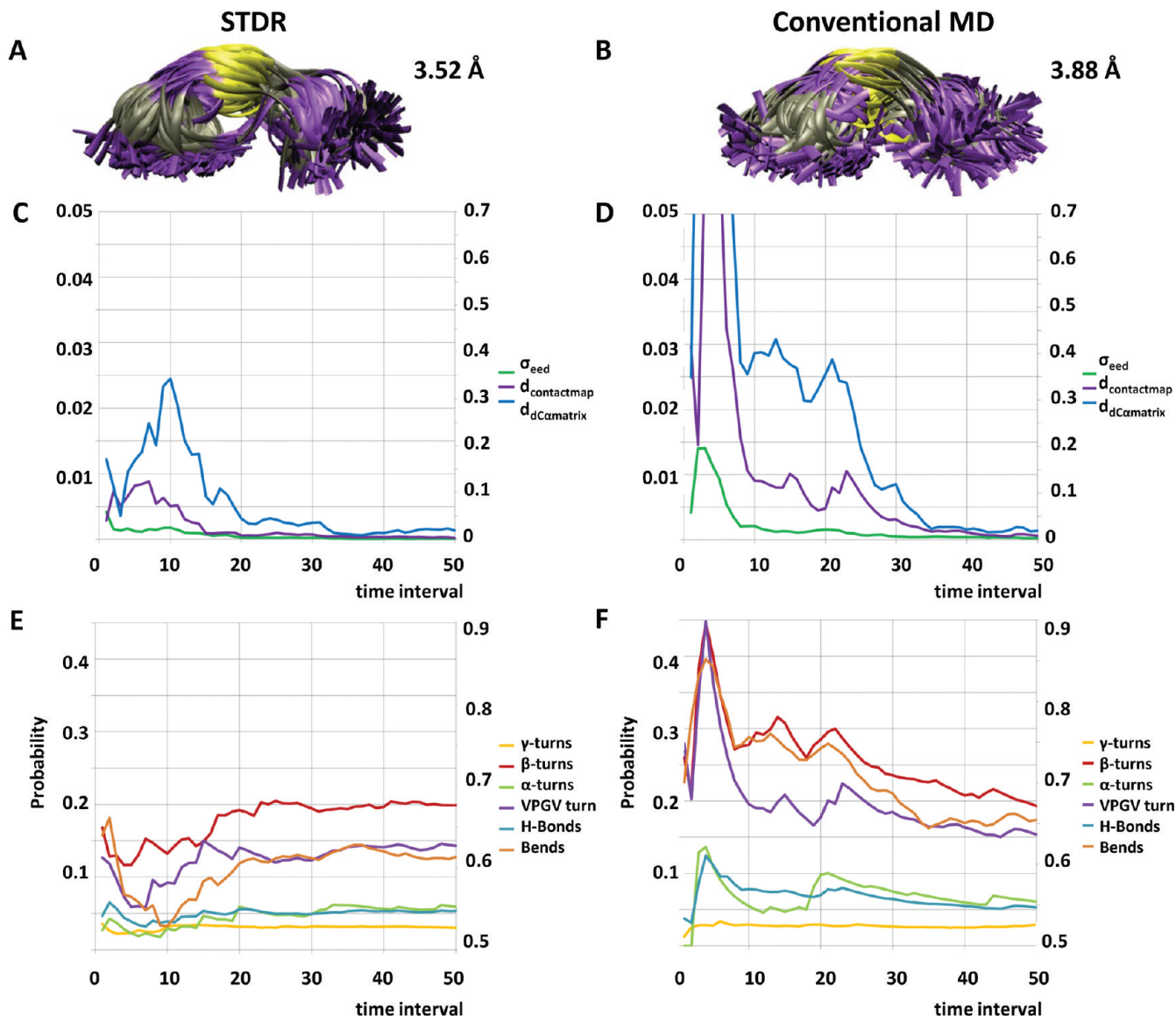
Figure 7. Comparing STDR and conventional MD for GVGVPGVG. (A) A total of 200 structures in ribbon representation obtained using the first 4.4 ns at 280 K for STDR are shown, and (B) for 144 ns of conventional MD, along with the corresponding rmsd. Glycine is in purple, proline is in yellow, and valine is in gray. C and D show σ_{eed} (eq 21), $d_{contactmap}$ (eq 23), and $d_{dC\alpha matrix}$ (eq 22, plotted on the secondary axis) for STDR and conventional MD, respectively. The trajectories are separated into 50 time intervals, and the quantities reported are calculated cumulatively, as in Figure 5. E and F show the probability per residue of a γ-turn, β-turn, and α-turn, as well as the population of the VPGV β-turn. The probabilities of observing an intramolecular hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown.

Choosing Between ST and STDR. For relatively small and simple systems, such as the octapeptide used in this study or a short polyalanine peptide in water,[30] the calculation of dimensionless Helmholtz free energies is possible, although computationally intensive. For these cases, ST is an ideal method, since it alleviates the need for communication between processors in parallel tempering and the subsequent waste of computational resources. However, calculation of the Helmholtz free energies increases in difficulty as system size and complexity increase. When the system is sufficiently large and complex, as is often the case for biomolecular systems of interest, limited computational resources may preclude the calculation of sufficiently accurate weight factors. That is, it is only possible to obtain dimensionless Helmholtz free energies which result in an acceptable level

of sampling uniformity with very extensive initial simulations. Even with near optimal weight factors updated throughout the simulation, Park and Pande still observed an average deviation from sampling homogeneity of 4.9% for a short peptide (calculated on the basis of the data in Table 1 of ref 30).[30] With very inaccurate weight factors, sampling of temperatures may be far from uniform. That is, there may be too little sampling at certain temperatures to obtain a reasonable estimate of the weight factors to facilitate adaptation. ST may therefore not be an appropriate method, even with adaptation of the weight factors throughout the simulation. This may be the case for many biomolecular systems of interest, which are larger than the small peptides or peptides in implicit solvent commonly used to test
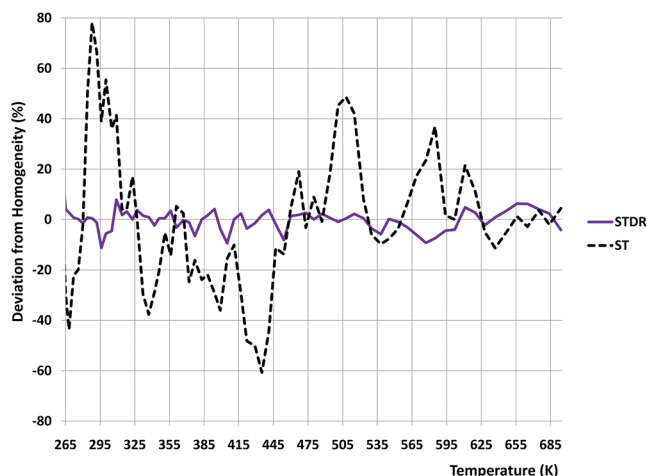
Figure 8. Deviation from sampling homogeneity for the ST simulation of $(GVPGV)_7$. For each temperature, the deviation from sampling homogeneity is computed as %deviation from homogeneity $= (N_m - \langle N_m \rangle)/\langle N_m \rangle \times 100\%$, where N_m is the number of samples at temperature T_m and $\langle N_m \rangle$ is the average number of samples per temperature. The ST simulation used the same weight factors as the STDR simulation, with 6 ns of sampling per replica (each started at a different temperature), for a total of 420 ns of simulation. A decrease in sampling between 410 and 460 K results in poor mobility of the replicas between high and low temperatures in the ST simulation.

generalized-ensemble methods. We now describe a complex system for which STDR is better suited than ST.

In addition to studying the octapeptide, GVGVPGVG, we also studied a longer peptide based on the same motif, $(GVPGV)_7$. Accurate weight factors for this system could not be obtained using a reasonable investment of computational resources (15 ns per temperature for 70 temperatures, for a total of 1.05 μs). Even with this large amount of data, the sampling of temperature in a ST simulation using these weight factors is heterogeneous. The average deviation from sampling homogeneity is 21.3% (computed using eq 20). In contrast, when STDR and the same weight factors are used, the average deviation from sampling homogeneity is only 3.4%. In Figure 8, the deviation from sampling homogeneity at each temperature is shown for both ST and STDR. In the ST simulation, sampling in the middle of the temperature range is less than that at both low and high temperatures, deviating from homogeneity by more than 60%. Since it is unfavorable in this case to sample intermediate temperatures, diffusion from high to low temperatures is impeded. In particular, the ST simulation using these weight factors experiences 56% fewer transitions between 417 and 454 K compared to the STDR simulation using the same weight factors. The sampling barrier in the intermediate temperature range impedes the random walk. The application of the DRPE results in a slight decrease of the acceptance ratio from 0.43 in the ST simulation to 0.38 in the STDR simulation. Importantly, the average replica speed in STDR is higher than that of the ST simulation (0.058 and 0.053 for STDR and ST, respectively). This indicates that replicas are able to efficiently explore temperature in the STDR simulation. In this case, the addition of the DRPE does not

significantly impair replica mobility and allows nearly uniform sampling of all temperatures. It is therefore a more suitable method than ST for this particular system.

We performed another comparison between ST and STDR using weight factors obtained with only 500 ps of canonical MD at each temperature. These weight factors are more inaccurate than those based on 15 ns of simulation at each temperature, leading to an average sampling inhomogeneity of 96.88% in a ST simulation. By applying the DRPE in a STDR simulation, the average sampling inhomogeneity is reduced to 8.84%. The acceptance ratio for the STDR simulation with these inaccurate weight factors is 0.38. This is exactly the same as the acceptance ratio of the STDR simulation using the weight factors based on 15 ns at each temperature. This observation is in agreement with the results of the STDR and STDRb simulations of the octapeptide. As shown in Table 2, both STDR and STDRb simulations exhibit nearly identical temperature diffusion coefficients, acceptance ratios, and mean free paths. Importantly, the results for both the octapeptide and the 35-residue peptide demonstrate that replica mobility in STDR simulations is not significantly affected by inaccuracy of the weight factors. Further, more accurate weight factors do not improve temperature diffusion, or corresponding structural convergence. Thus, we recommend that, when the STDR algorithm is used, the computational investment for the initial calculation of weight factors should be minimized.

Performance of STDR for More Complex Systems. To compare STDR with conventional MD, we also performed a MD simulation at the lowest temperature of the STDR simulation (261 K). A superposition of 200 structures, obtained every 1 ns from a 200 ns trajectory generated using conventional MD at 261 K, is shown in Figure 9A. These structures have an average rmsd of 1.66 Å, indicating that the peptide is trapped in one conformational basin and undergoes only small conformational changes. This set of structures contrasts with the set of 200 randomly selected structures from the complete STDR simulation at 261 K (Figure 9B), which represents completely different conformations with an average rmsd of 8.40 Å. For clarity, we also show six example structures in Figure 9D to demonstrate the variety of conformations observed in the STDR simulation. To make a more direct comparison between conventional MD and STDR, Figure 9C shows 200 structures from STDR using the same amount of simulation time as the conventional MD simulation (200 ns summed over all of the temperatures, corresponding to approximately 3 ns at 261 K). Finally, the radius of gyration distributions for conventional MD, STDR, and the first 3 ns of STDR are shown in Figure 9E. Both distributions from STDR show several conformational states, while the conventional MD simulation is trapped in one state. Even when using the same amount of computational resources, STDR produces a more heterogeneous ensemble of conformations.

Figure 10 displays hydrogen-bonding contact maps at 261 K for STDR, STDR with 3 ns of sampling, and conventional MD. STDR produces a conformational ensemble in which many contacts are formed with low probability. In contrast, conventional MD generates a contact map with only a few
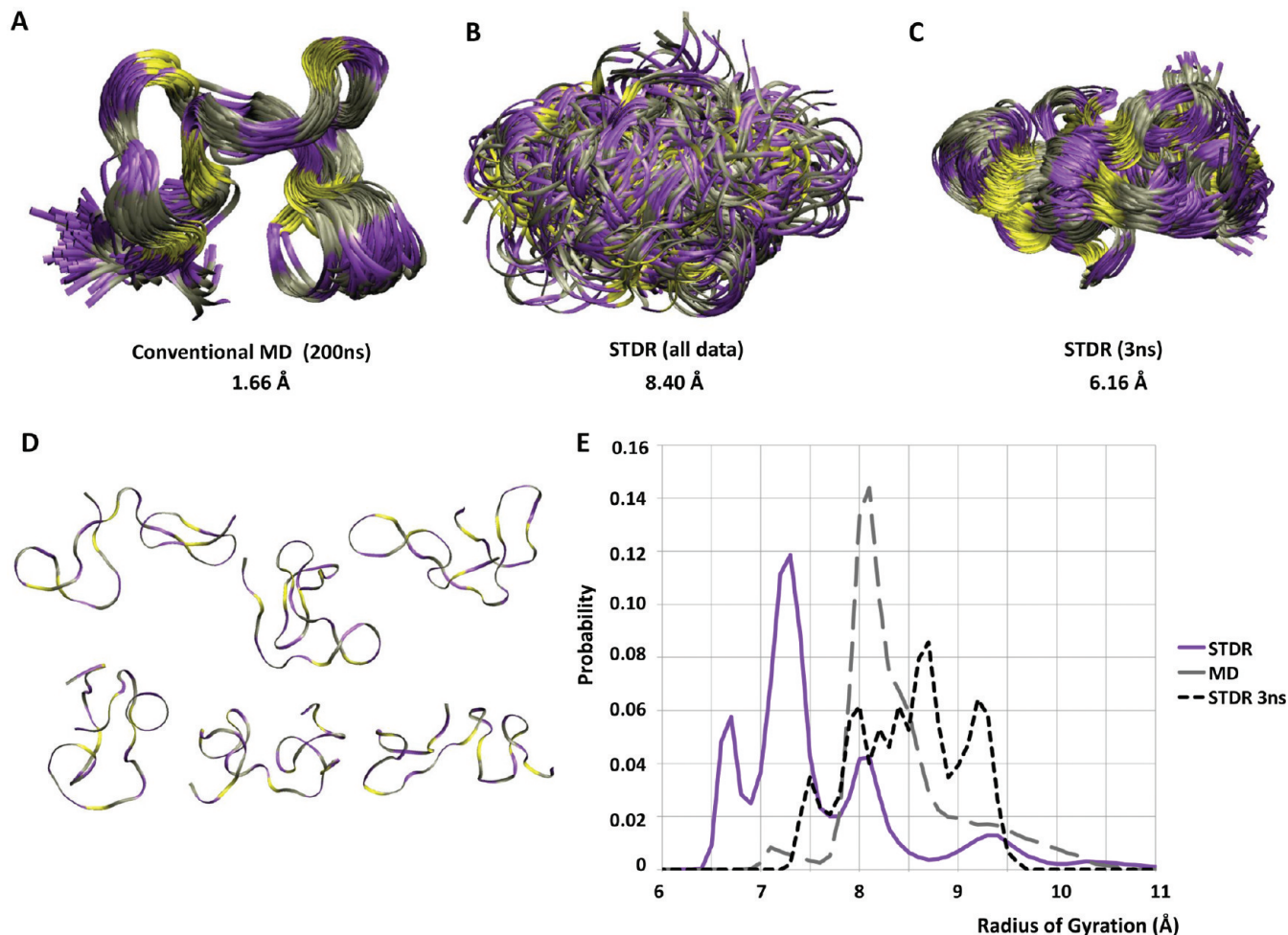
Figure 9. Comparing STDR and conventional MD for (GVPGV)$_7$. A total of 200 structures in ribbon representation along with their rmsd are shown for (A) the conventional MD simulation of length 200 ns at 261 K, (B) for the STDR simulation at 261 K using all of the data (120 ns at this temperature), (C) and for the STDR simulation at 261 K using the first 3 ns (this is the same simulation time summed over all replicas as A). Glycine is in purple, proline is in yellow, and valine is in gray. (D) A selection of six example structures is shown from the structures in B to illustrate the structural diversity obtained using STDR. (E) Probability distributions of the radius of gyration.

contacts, some of which are formed for nearly the entire simulation. The contact maps are shown with two different vertical scales to emphasize this point. If only the conventional MD simulation had been performed, a completely different understanding of the conformational landscape would have emerged. Single-temperature MD severely underestimates the heterogeneity of the conformational landscape and exhibits both pseudoconvergence and quasi-nonergodicity. Even when using the same amount of simulation time as MD, the contact map from STDR has more contacts, none of which has a probability of more than 30%.

It is not possible to quantitatively measure the speedup of STDR versus conventional MD for this system because limited computational resources preclude performing conventional MD simulations for the time required to achieve structural convergence. We observe that conventional MD is trapped in one conformational basin for 200 ns. It is not possible to accurately predict how long it would take to sample all relevant states and reach convergence. Qualitatively, we observe a dramatic sampling enhancement due to STDR. Using the same amount of computational resources, STDR generates more unique conformations for this peptide.

This indicates that the random walk in temperature does in fact lead to enhanced sampling, establishing the efficacy of the STDR method for a complex polypeptide.

Before sampling the complete energy landscape of a system of interest, there is no way to confidently predict the height of the energy barriers, or the number of energetically stable conformations (local minima of the energy surface). By coupling to simulations at higher temperatures, high energetic barriers can be overcome. However, if one is not simultaneously interested in the behavior of the system at multiple temperatures, it may be less computationally expensive to run very long simulations, or a collection of simulations, at a single temperature.[75] In the present study, conventional MD successfully produced the conformational ensemble of the octapeptide but resulted in quasi-nonergodicity for the 35-residue peptide. In order to truly "enhance sampling" relative to single-temperature MD simulations, an enhanced sampling method must achieve convergence at a rate which is greater than the product of the number of replicas and the computer time for each replica.

The present study of (GVPGV)$_7$ shows that it is possible to observe pseudoconvergence using single-temperature MD

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2659**
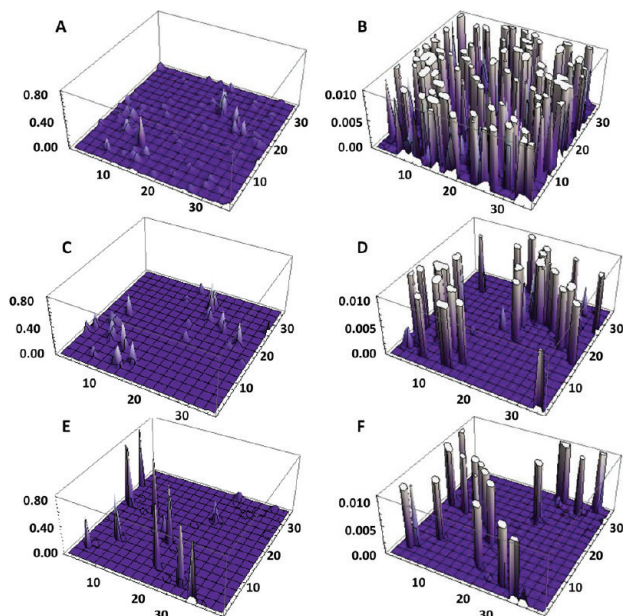
Figure 10. Hydrogen-bonding contact maps from STDR and conventional MD. Hydrogen-bonding contact maps are depicted as three-dimensional maps, where peak height represents the probability of contact formation. These plots are shown on two scales. On the left, the scale has a maximum of 0.8, and on the right, 0.01, for clarity in showing the contacts formed with low probability. A and B, the STDR simulation at 261 K using all of the data (120 ns at this temperature). C and D, the STDR simulation using the first 3 ns (this is the same simulation time summed over all replicas as E and F). E and F, 200 ns of conventional MD. Some contacts are formed over 80% of the time.

(that is, to observe convergence of a quantity of interest without observing the true value of that quantity, Boltzmann-weighted by the populations of all possible conformations). Long-time MD simulations do not yield the appropriate conformational distribution, and the system remains trapped in a local minimum of the energy landscape. In contrast, we observe that conventional MD is able to satisfactorily reproduce the conformational ensemble of the octapeptide at a significantly reduced computational cost compared to using a generalized-ensemble method. In light of this apparent contradiction, how are the averages of quantities obtained using MD simulations to be interpreted? On the basis of this work, it appears that using generalized-ensemble algorithms is a more prudent approach, even if in some cases it may be less efficient overall to do so (for increased confidence in the accuracy of the data). Several other examples have shown that the enhanced sampling provided by generalized-ensemble methods provides convergence that would not be feasible with single-temperature MD.[6,14,18−24] These observations underscore not only the need for enhanced sampling methods but also the shortcomings of techniques such as block averaging over simulations initiated in a single conformational basin in estimating the convergence of results. The challenge in simulating complex systems is that a priori one does not know the efficiency of the generalized-ensemble approach relative to the "brute force" MD approach. It may be advisable to use a general-

ized-ensemble algorithm, especially if conformational sampling, and not dynamic information, is sought.

Conclusions

We now return to the original question: given limited computational resources, which generalized-ensemble algorithm is most efficient at sampling a complex conformational landscape? The first important distinction between methods is the separation between those based on ST and those based on RE. In this paper, we demonstrate that ST-based methods result in both faster temperature diffusion and faster structural convergence. They are therefore preferable to RE-based methods. This is the most general conclusion of the paper.

Within the family of RE-based methods, the computational efficiencies of the various algorithms are not equivalent. SREM should only be applied to systems for which PEDFs can be accurately obtained. Therefore, due to limited computational resources, SREM can only be applied to simple systems. Like SREM, RE is not well-suited to complex systems because of the need to synchronize simulations of a large number of replicas (and typically, a large number of processors). Although there is no theoretical limit on the number of replicas that one can use for a RE simulation, it is generally difficult in practice to obtain access to a large, dedicated, and homogeneous computing cluster. Even if one does have access to such a computational resource, the wasted CPU time may also increase sharply with the number of replicas due to both CPU failure and inhomogeneity in CPU speeds.

VREX represents an attractive alternative to RE since it completely eliminates synchronization and communication between replicas. It produces more homogeneous sampling of temperature compared to SREM, with much less initial simulation time. It is therefore preferable to both SREM and RE. Since RE-based methods suffer from slower structural convergence and temperature diffusion compared to ST-based methods, it is preferable to use a ST-based method in temperature. This disadvantage may become less significant for complex systems. VREX may also be a more suitable method in another reaction coordinate other than temperature, for which weight factors are much more difficult to obtain. Moreover, initial simulations for weight factor calculations may also benefit from the use of VREX.

In the case of relatively simple systems for which weight factors can be accurately calculated using minimal computational resources, ST is the most appropriate method. We have shown that ST with accurate weight factors exhibits the fastest temperature diffusion and, correspondingly, the fastest structural convergence. However, for more complex systems, for which weight factors are costly to compute, STDR becomes the preferred method. Even with infinite resources, a long initial simulation to compute weight factors accurate enough to yield homogeneous sampling in ST may not be the most efficient use of computational resources. The feasibility of a ST simulation is determined by the accuracy of the weight factors, which can only be assessed by actually performing a ST simulation. Importantly, we have demonstrated that STDR can make use of inaccurate weight factors to achieve homogeneous sampling of temperature and

consequently structural convergence. Replica mobility is only slightly impeded by the DRPE. By contrast, ST with inaccurate weight factors produces heterogeneous sampling of temperatures, which is also an impediment to the random walk in temperature. STDR is suitable for any computing cluster or distributed computing environment, since it requires no fixed number of CPUs or synchronization of exchanges. Complex systems can therefore benefit from the STDR method, which is increasingly advantageous as system size and complexity grow.

Frequently Used Abbreviations

CPU, central processing unit; DR, distributed replica sampling; DRPE, distributed replica potential energy; EED, end-to-end distance; MD, molecular dynamics; RE, replica exchange; SREM, serial replica exchange method; ST, simulated tempering; STDR, simulated tempering distributed replica sampling; VREX, virtual replica exchange.

Appendix 1. Computing Error in Exchange Probabilities for ST and SREM

The PEDFs of the octapeptide are nearly perfect Gaussian distributions, as expected due to the large number of degrees of freedom of the system and the central limit theorem.[14] Assuming that the PEDFs are Gaussian is in general a valid assumption for biomolecular systems.[76] As an estimate of the error in the PEDFs, we considered the average deviation of the average energy of each PEDF, $\langle E_n \rangle$, from the average energy of a reference PEDF, $\langle E_n \rangle_{\text{reference}}$, as follows:

$$\sigma_{\text{PEDFs}} \approx \frac{1}{N_{\text{temps}}} \sum_{n=1}^{N_{\text{temps}}} |\langle E_n \rangle - \langle E_n \rangle_{\text{reference}}| \quad (24)$$

where N_{temps} is the number of temperatures. For the reference PEDFs, we used potential energy distribution functions calculated on the basis of all of the data from the RE simulation (a total of 4.75 μs for all temperatures). We computed the average error in the differences of weight factors in an analogous way, also using the RE simulation as reference data:

$$\sigma_{\text{weightfactors}} \approx \frac{1}{N_{\text{temps}} - 1} \sum_{n=1}^{N_{\text{temps}}-1} |(a_{n+1} - a_n) - (a_{n+1} - a_n)_{\text{reference}}| \quad (25)$$

The selection of the RE simulation as a reference was made because it was the only generalized-ensemble method that we tested that did not make use of any initial simulation.

In order to make a fair comparison between the errors in the weight factors used in ST and the PEDFs used in SREM, it is important to consider the error in not only the potential energy distribution functions and dimensionless Helmholtz free energies but also the error in the resulting exchange probabilities. The error in the exchange probability of SREM (eq 9) was computed as follows:

$$\sigma_{P_{ij}} = \sqrt{\left[\left(\frac{\partial P_{ij}}{\partial E_{j,\text{PEDF}}}\right)\sigma_{E_{j,\text{PEDF}}}\right]^2} = e^{-(\beta_j - \beta_i)(E_i - E_{j,\text{PEDF}})}(\beta_j - \beta_i)\sigma_{E_{j,\text{PEDF}}} \quad (26)$$

We estimate this error by using the estimate for the error in the PEDFs obtained in eq 24 and the average acceptance ratio and average difference in inverse temperatures:

$$\sigma_{P_{ij},\text{estimate}} \approx \langle e^{-(\beta_j - \beta_i)(E_i - E_{j,\text{PEDF}})} \rangle \langle \beta_j - \beta_i \rangle \sigma_{\text{PEDFs}} \quad (27)$$

Similarly, the error in the exchange probability for ST (given by eq 6) is

$$\sigma_{P_{ij}} = \sqrt{\left[\left(\frac{\partial P_{ij}}{\partial (a_j - a_i)}\right)\sigma_{(a_j - a_i)}\right]^2} = e^{(\beta_j - \beta_i)E - (a_j - a_i)}\sigma_{(a_j - a_i)} \quad (28)$$

and this error is estimated using the average error in the weight factor differences from eq 25 and the average acceptance ratio:

$$\sigma_{P_{ij},\text{estimate}} \approx \langle e^{(\beta_j - \beta_i)E - (a_j - a_i)} \rangle \sigma_{\text{weightfactors}} \quad (29)$$

Supporting Information Available: Supplementary table (S1), supplementary figures (S1−S3), and an example calculation of the DRPE are provided. This material is available free of charge via the Internet at http://pubs.acs.org. Distributed Replica Sampling software is available online at www.pomeslab.com.

References

(1) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; García, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13* (2), 168–174.

(2) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60* (2), 96–123.

(3) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19* (6), 451–458.

(4) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96* (3), 1776–1783.

(5) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140–150.

(6) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141–151.

(7) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65* (6), 1604–1608.

(8) Tesi, M. C.; van Rensburg, E. J. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82* (1−2), 155–181.

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2661**

(9) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61* (23), 2635–2638.

(10) Hagen, M.; Kim, B.; Liu, P.; Friesner, R. A.; Berne, B. J. *J. Phys. Chem. B* **2007**, *111* (6), 1416–1423.

(11) Rodinger, T.; Howell, P. L.; Pomès, R. *J. Chem. Theory Comput.* **2006**, *2* (3), 725–731.

(12) Rodinger, T.; Howell, P. L.; Pomès, R. *J. Chem. Phys.* **2008**, *129* (15), 155102.

(13) Neale, C.; Rodinger, T.; Pomès, R. *Chem. Phys. Lett.* **2008**, *460* (1−3), 375–381.

(14) Huang, X.; Bowman, G. R.; Pande, V. S. *J. Chem. Phys.* **2008**, *128* (20), 205106.

(15) Chipot, C.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer: Berlin, 2007.

(16) Bedrov, D.; Smith, G. D. *J. Chem. Phys.* **2001**, *115* (3), 1121–1124.

(17) Yamamoto, R.; Kob, W. *Phys. Rev. E* **2000**, *61* (5), 5473–5476.

(18) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46* (2), 225–234.

(19) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123* (15), 154105.

(20) Nymeyer, H. *J. Chem. Theory Comput.* **2008**, *4* (4), 626–636.

(21) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126* (1), 014903.

(22) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119* (7), 4035–4042.

(23) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84* (2), 775–786.

(24) Tsai, H. H.; Reches, M.; Tsai, C. J.; Gunasekaran, K.; Gazit, E.; Nussinov, R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (23), 8174–8179.

(25) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458* (1−3), 244–248.

(26) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2* (4), 1200–1202.

(27) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2* (6), 1693–1693.

(28) Beck, D. A. C.; White, G. W. N.; Daggett, V. *J. Struct. Biol.* **2007**, *157* (3), 514–523.

(29) Hansmann, U. H. E.; Okamoto, Y. *Phys. Rev. E* **1996**, *54* (5), 5863–5865.

(30) Park, S.; Pande, V. S. *Phys. Rev. E* **2007**, *76* (1), 016703.

(31) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425–439.

(32) Zhang, C.; Ma, J. P. *J. Chem. Phys.* **2008**, *129* (13), 7.

(33) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1997**, *18* (7), 920–933.

(34) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021.

(35) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3* (1), 26–41.

(36) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332* (1−2), 131–138.

(37) Park, S. *Phys. Rev. E* **2008**, *77* (1), 016709.

(38) Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2004**, *121* (6), 2491–2504.

(39) Zhang, C.; Ma, J. P. *Phys. Rev. E* **2007**, *76* (3), 036708.

(40) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86* (10), 2050–2053.

(41) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1993**, *14* (11), 1333–1338.

(42) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116* (20), 9058–9067.

(43) Gallicchio, E.; Levy, R. M.; Parashar, M. *J. Comput. Chem.* **2008**, *29* (5), 788–794.

(44) Luckow, A.; Jha, S.; Kim, J.; Merzky, A.; Schnor, B. *Phil. Trans. R. Soc. A* **2009**, *367* (1897), 2595–2606.

(45) Woods, C. J.; Ng, M. H.; Johnston, S.; Murdock, S. E.; Wu, B.; Tai, K.; Fangohr, H.; Jeffreys, P.; Cox, S.; Frey, J. G.; Sansom, M. S. P.; Essex, J. W. *Phil. Trans. R. Soc. A* **2005**, *363* (1833), 2017–2035.

(46) Shirts, M.; Pande, V. S. *Science* **2000**, *290* (5498), 1903–1904.

(47) Bowman, G. R.; Huang, X.; Yao, Y.; Sun, J.; Carlsson, G.; Guibas, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2008**, *130* (30), 9676–9678.

(48) Shen, H.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2008**, *4* (8), 1386–1400.

(49) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2009**, *5* (3), 627–640.

(50) Henry, R. M.; Yu, C. H.; Rodinger, T.; Pomès, R. *J. Mol. Biol.* **2009**, *387* (5), 1165–1185.

(51) Miao, M.; Bellingham, C. M.; Stahl, R. J.; Sitarz, E. E.; Lane, C. J.; Keeley, F. W. *J. Biol. Chem.* **2003**, *278* (49), 48553–48562.

(52) Rauscher, S.; Baud, S.; Miao, M.; Keeley, F. W.; Pomès, R. *Structure* **2006**, *14* (11), 1667–1676.

(53) Baer, M.; Schreiner, E.; Kohlmeyer, A.; Rousseau, R.; Marx, D. *J. Phys. Chem. B* **2006**, *110* (8), 3576–3587.

(54) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(55) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7* (8), 306–317.

(56) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(57) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(59) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(60) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(61) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(62) Nosé, S. *Mol. Phys.* **1984**, *52* (2), 255–268.

(63) Hoover, W. G. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

(64) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21* (6), 1087–1092.

(65) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

(66) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.

(67) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.

(68) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22* (12), 2577–2637.

(69) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14* (1), 33–38.

(70) Abraham, M. J.; Gready, J. E. *J. Chem. Theory Comput.* **2008**, *4* (7), 1119–1128.

(71) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (39), 13749–13754.

(72) Thirumalai, D.; Mountain, R. D.; Kirkpatrick, T. R. *Phys. Rev. A* **1989**, *39* (7), 3563–3574.

(73) Rousseau, R.; Schreiner, E.; Kohlmeyer, A.; Marx, D. *Biophys. J.* **2004**, *86* (3), 1393–1407.

(74) Schreiner, E.; Nicolini, C.; Ludolph, B.; Ravindra, R.; Otte, N.; Kohlmeyer, A.; Rousseau, R.; Winter, R.; Marx, D. *Phys. Rev. Lett.* **2004**, *92* (14), 148101.

(75) Monticelli, L.; Sorin, E. J.; Tieleman, D. P.; Pande, V. S.; Colombo, G. *J. Comput. Chem.* **2008**, *29* (11), 1740–1752.

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

JCTC Journal of Chemical Theory and Computation

Exchange-Dispersion Energy: A Formulation in Terms of Monomer Properties and Coupled Cluster Treatment of Intramonomer Correlation

Tatiana Korona*

Faculty of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland

Received May 9, 2009

Abstract: A new formulation of the second-order exchange-dispersion energy of symmetry-adapted perturbation theory is presented. The proposed approach allows to study − for the first time for many-electron monomers − the effect of intramonomer electron correlation on this energy correction. In the new formalism the exchange-dispersion energy is expressed in terms of properties of interacting molecules, such as dynamic density-matrix susceptibilities and one-electron reduced density matrices. The resulting formula has been used to obtain the main (noncumulant) part of the exchange-dispersion energy for monomers described by coupled cluster theory limited to single and double excitations. A density fitting approach has been applied in order to reduce the computational effort for obtaining coupled cluster density-matrix susceptibilities. The new coupled cluster exchange-dispersion energy has been compared with other available values of this interaction energy component, obtained with monomers treated on the level of Hartree−Fock or density-functional theories.

1. Introduction

Symmetry-adapted perturbation theory (SAPT) (see refs 1–3 for reviews) is a well established alternative to the super-molecular approach for calculation of interaction energies between two closed-shell molecules A and B (called monomers in the following). In SAPT, one divides the total Hamiltonian H of a dimer AB into the unperturbed part H_0, being the sum of monomers' Hamiltonians H_A and H_B, and a perturbation V, which describes the electrostatic interaction between monomers (N_A and N_B denote numbers of electrons in the monomers A and B, respectively):

$$V = \sum_{i=1}^{N_A} \sum_{j=1+N_A}^{N_A+N_B} \frac{1}{r_{ij}} - \sum_{i=1}^{N_A} \sum_{\beta \in B} \frac{Z_\beta}{r_{\beta i}} - \sum_{j=1+N_A}^{N_A+N_B} \sum_{\alpha \in A} \frac{Z_\alpha}{r_{\alpha j}} + \sum_{\alpha \in A} \sum_{\beta \in B} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \quad (1)$$

The intermolecular interaction operator V can be conveniently rewritten in the form:

$$V = \sum_{i=1}^{N_A} \sum_{j=1+N_A}^{N_A+N_B} v(i,j) \quad (2)$$

where a generalized interaction operator $v(i, j)$ is defined as[4]

$$v(i,j) = \frac{1}{r_{ij}} - \frac{1}{N_B} \sum_{\beta \in B} \frac{Z_\beta}{r_{\beta i}} - \frac{1}{N_A} \sum_{\alpha \in A} \frac{Z_\alpha}{r_{\alpha j}} + \frac{1}{N_A N_B} \sum_{\alpha \in A} \sum_{\beta \in B} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \quad (3)$$

[In the following the coordinates (including spin) of the ith electron will be denoted by a number i, while spatial coordinates of this electron will be defined as $\mathbf{r}_i$.]

In SAPT, the interaction energy up to the second order in V is defined as a sum of three Rayleigh−Schrödinger corrections: first-order electrostatic ($E_{elst}^{(1)}$), second-order induction ($E_{ind}^{(2)}$) and second-order dispersion ($E_{disp}^{(2)}$) energies and their exchange counterparts, accounting for electron tunneling effects between monomers: first-order exchange ($E_{exch}^{(1)}$), second-order exchange-induction ($E_{exch-ind}^{(2)}$), and second-order exchange-dispersion ($E_{exch-disp}^{(2)}$) energies. In textbook derivations of perturbation theories it is always

* Corresponding author. E-mail: tatiana.korona@chem.uw.edu.pl.

Table 1. The MP Level of SAPT(MP) Corrections (As Derived and Programmed in SAPT Program[13])[a]

energy	MP level	notes
$E_{elst}^{(1)}$	3	relaxed (preferred) and unrelaxed versions available. The fourth order is also present, although not used in practice
$E_{ind}^{(2)}$	2	unrelaxed version (*true* $^t E_{ind}^{(22)}$),[2,10] the relaxed version is available in the literature but absent in the SAPT program
$E_{disp}^{(2)}$	2	CCD+ST(CCD)[12] dispersion energy is also available in SAPT; the coupled Hartree−Fock (CHF) dispersion energy also exists in the literature[18]
$E_{exch}^{(1)}$	2	usually CCSD monomer amplitudes are used instead of MP1 and MP2 ones in formulas for the $E_{exch}^{(1ij)}$ energy components[11]
$E_{exch-ind}^{(2)}$	0	relaxed (preferred) and unrelaxed versions available, the scaling formula is used to estimate the intramonomer correlation effects
$E_{exch-disp}^{(2)}$	0	unrelaxed version

[a] Some corrections are available in relaxed and unrelaxed versions, depending on whether monomer orbitals are or are not allowed to respond to the perturbation field of another monomer.

assumed that eigenvalues and eigenvectors of the H_0 part of the total Hamiltonian H are easy to calculate. However, it is well-known that the exact solutions of the zeroth-order Hamiltonian $H_0 = H_A + H_B$ are not available (with the exception of the smallest few-electron molecules). Therefore, additional approximations are necessary in order to obtain SAPT interaction energies for many-electron monomers.

Since usually the eigenfunctions of the Fock operator can be easily obtained, Hartree−Fock (HF) determinants are commonly used as a zeroth-order approximation to the unknown exact wave functions of monomers, giving rise to the SAPT(HF) approach. [A convention will be adopted in this paper to denote as SAPT(Z) the SAPT approach with monomers treated in the level Z.] Obviously, SAPT(HF) neglects the effect of Coulomb electron correlation inside the monomers. In order to account for this effect, every SAPT correction is expanded in terms of monomer fluctuation operators $W_X = H_X - F_X$, X = A, B, leading to the triple perturbation theory.[5] Effectively, in this approach the Møller−Plesset (MP) expansion is utilized, so the method can be labeled as SAPT(MP).[4,6–12] Energy corrections in SAPT(MP) are denoted as $E^{(nij)}$, where n, i and j are orders of the perturbation operators V, W_A, and W_B, respectively. The SAPT(MP) approach has been implemented into the SAPT suite of codes[13] and has proven to be an efficient method for studying noncovalent complexes of small atoms and molecules (for examples of applications see refs 14–17). Levels of MP theory used for various SAPT corrections are presented in Table 1. In particular, it can be seen from this table that the exchange-dispersion energy in SAPT(MP) has been so far treated on the uncorrelated and uncoupled level (i.e., neither the electron correlation nor the orbital relaxation effects are taken into account for this correction). A lack of

the intramonomer correlation for the exchange-dispersion energy is potentially one of the main sources of errors in SAPT(MP).

Yet another treatment of the intramonomer correlation problem is based on a density-functional theory (DFT) description of monomers.[19,20] Two groups were involved simultaneously in a development of SAPT(DFT): Misquitta, Szalewicz, and Jeziorski[21,22] (see also ref 23) and Jansen and Hesselmann[24–27] (the latter authors prefer to use the abbreviation DFT-SAPT for their method). The implementation of SAPT(DFT), followed by the utilization of the density fitting (DF) approximation[28] to electron-repulsion integrals,[29–32] allows nowadays the use of SAPT for large van der Waals molecules, like dimers of cyclotrimethylene trinitramine[33] and DNA bases[34] or even for endohedral complexes of fullerenes.[35]

Very recently a new SAPT(CC) approach has been proposed.[36–42] In this method the wave functions of monomers are described by coupled cluster (CC) theory.[43] So far the CC treatment has been applied to the electrostatic,[36,37] first-order exchange,[40,41] induction,[38,42] exchange-induction,[42] and dispersion[39] terms. A new formalism proposed in this paper enables the calculation of the second-order exchange-dispersion energy with an account of intramonomer correlation. In practice, the method has been implemented for the case of monomers described by CC theory limited to single and double excitations (CCSD).

The plan for this paper is the following: first the exchange-dispersion energy will be expressed in terms of monomer properties in an explicitly connected form. Next the orbital working formula will be derived, and a DF approximation will be applied. Finally, the illustrative results for the CCSD exchange-dispersion energy will be presented and discussed.

2. Theory

2.1. Exchange-Dispersion Energy through Monomer Properties.
The main goals of the derivations in the next sections are to describe the exchange-dispersion energy in terms of monomer properties and to express the resulting formula in a computationally convenient form.

It is known that if monomers are not too close to each other, then one can neglect multiple exchanges of electrons between them, what leads to the so-called single-exchange or S^2 approximation.[44] The second-order exchange-dispersion energy in the S^2 approximation is given by the following formula:[45,46]

Exchange-Dispersion Energy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2665**

$$E_{\text{exch-disp}}^{(2)} = \langle \Psi_A^0 \Psi_B^0 | VP\Psi_{\text{disp}}^{(1)} \rangle - \bar{V} \langle \Psi_A^0 \Psi_B^0 | P\Psi_{\text{disp}}^{(1)} \rangle - \bar{P} \langle \Psi_A^0 \Psi_B^0 | V\Psi_{\text{disp}}^{(1)} \rangle \tag{4}$$

In eq 4, Ψ_A^0 and Ψ_B^0 are the exact wave functions for the ground states of the monomers A and B, and P denotes the single-exchange operator:

$$P = -\sum_{i=1}^{N_A} \sum_{j=1+N_A}^{N_A+N_B} P_{ij} \tag{5}$$

while the first-order dispersion wave function is obtained from the following equation:

$$\Psi_{\text{disp}}^{(1)} = -\sum_{K=1}^{\infty} \sum_{L=1}^{\infty} \frac{|\Psi_A^K \Psi_B^L\rangle\langle\Psi_A^K \Psi_B^L | V_{ee} \Psi_A^0 \Psi_B^0\rangle}{\Delta E_A^K + \Delta E_B^L} \tag{6}$$

where indices $K(L)$ number the eigenstates of the monomer A(B), ΔE_A^K and ΔE_B^L denote excitation energies of the monomers A and B from their ground states, and V_{ee} is a pure two-electron part of the operator V. Finally, the symbol $\bar{X}$ denotes the mean value of an operator X with the zeroth-order wave function, i.e., with $\Psi_A^0 \Psi_B^0$ (so, e.g. $\bar{V} = E_{\text{elst}}^{(1)}$).

Let us consider the term:

$$-\langle \Psi_A^0 \Psi_B^0 | VP\Psi_A^K \Psi_B^L \rangle = \int \Psi_A^0(1, 2, ..., N_A)^* \Psi_B^0(1', 2', ..., N_B')^* \sum_{i=1}^{N_A} \sum_{i'=1'}^{N_B'} v(i, i') \times$$
$$\sum_{j=1}^{N_A} \sum_{j'=1'}^{N_B'} P_{jj'} \Psi_A^K(1, 2, ..., N_A) \Psi_B^L(1', 2', ..., N_B') d\tau_1 ... d\tau_{N_B'} \tag{7}$$

which is obtained when eq 6 is inserted into eq 4 (electrons of the monomer B are denoted by primes, i.e., $i' = i + N_A$). After making use of the indistinguishability of electrons within the monomers the following four types of integrals are obtained from eq 7:

$$\begin{aligned}
-\langle \Psi_A^0 \Psi_B^0 | VP\Psi_A^K \Psi_B^L \rangle = {} & N_A N_B \int \Psi_A^0(1, 2, ..., N_A)^* \Psi_B^0(1', 2', ..., N_B')^* v(1, 1') \\
& \times \; [\Psi_A^K(1', 2, ..., N_A)\Psi_B^L(1, 2', ..., N_B') \\
& + (N_B - 1)\Psi_A^K(2', 2, ..., N_A)\Psi_B^L(1', 1, ..., N_B') \\
& + (N_A - 1)\Psi_A^K(1, 1', ..., N_A)\Psi_B^L(2, 2', ..., N_B') \\
& + (N_A - 1)(N_B - 1)\Psi_A^K(1, 2', ..., N_A)\Psi_B^L(1', 2, ..., N_B')] d\tau_1 ... d\tau_{N_B'}
\end{aligned} \tag{8}$$

In eq 8, the reduced one- and two-electron transition density matrices from the ground to the Kth (Lth) excited state of the monomer A(B) can be identified. For the monomer A, these transition density matrices are given by the formulas,

$$\begin{aligned}
\rho_A^{0 \to K}(1|1') &= N_A \int \Psi_A^0(1, 2, ..., N_A)^* \Psi_A^K(1', 2, ..., N_A) d\tau_2 ... d\tau_{N_A} \\
\Gamma_A^{0 \to K}(12|1'2') &= N_A(N_A - 1) \int \Psi_A^0(1, 2, ..., N_A)^* \Psi_A^K(1', 2', ..., N_A) d\tau_3 ... d\tau_{N_A}
\end{aligned} \tag{9}$$

Note that the ground-state density matrices $\rho_A(1|1')$ and $\Gamma_A(12|1'2')$ can be obtained by replacing Ψ_A^K by Ψ_A^0 in eq 9. The definitions from eq 9 allow us to rewrite the term under study in the following way:

$$\begin{aligned}
-\langle \Psi_A^0 \Psi_B^0 | VP\Psi_A^K \Psi_B^L \rangle = {} & \int \rho_A^{0 \to K}(1|1')\rho_B^{0 \to L}(1'|1)v(1, 1')d\tau_1 d\tau_1' \\
& + \int \rho_A^{0 \to K}(1|2')\Gamma_B^{0 \to L}(1'2'|1'1)v(1, 1')d\tau_1 d\tau_1' d\tau_2' \\
& + \int \Gamma_A^{0 \to K}(12|11')\rho_B^{0 \to L}(1'|2)v(1, 1')d\tau_1 d\tau_1' d\tau_2 \\
& + \int \Gamma_A^{0 \to K}(12|12')\Gamma_B^{0 \to L}(1'2'|1'2)v(1, 1')d\tau_1 d\tau_1' d\tau_2 d\tau_2'
\end{aligned} \tag{10}$$

An analogous derivation for the integral present in the numerator of eq 6 leads to a similar result,

$$\langle \Psi_A^K \Psi_B^L | V_{ee} \Psi_A^0 \Psi_B^0 \rangle = \int \rho_A^{K \to 0}(1|1)\rho_B^{L \to 0}(1'|1')\frac{1}{r_{11'}}d\tau_1 d\tau_1' \tag{11}$$

The final step needed to rewrite eq 4 through quantities belonging to the monomers A and B, is performed by utilizing the equation for energy denominators:

$$\frac{1}{\Delta E_{A}^{K} + \Delta E_{B}^{L}} = \frac{2}{\pi} \int_{0}^{\infty} \frac{\Delta E_{A}^{K}}{(\Delta E_{A}^{K})^{2} + \omega^{2}} \frac{\Delta E_{B}^{L}}{(\Delta E_{B}^{L})^{2} + \omega^{2}} d\omega \tag{12}$$

used by Longuet-Higgins[47] to derive his famous formula for the second-order dispersion energy in terms of frequency-dependent density susceptibilities of the monomers A and B, $\alpha_{X}(\mathbf{r},\mathbf{r}'|\omega)$:

$$E_{\text{disp}}^{(2)} = -\frac{1}{2\pi} \int_{0}^{\infty} \int \alpha_{A}(\mathbf{r}_{1}, \mathbf{r}_{2}|i\omega)\alpha_{B}(\mathbf{r}_{3}, \mathbf{r}_{4}|i\omega)\frac{1}{r_{13}}\frac{1}{r_{24}}d\mathbf{r}_{1}d\mathbf{r}_{2}d\mathbf{r}_{3}d\mathbf{r}_{4}d\omega \tag{13}$$

By combining eqs 4, 10, 11, and 12, one expresses the exchange-dispersion energy in terms of monomer properties, at the cost of making an integration over the frequency ω. The monomer properties, which have been identified in this process, are straightforward generalizations of frequency-dependent density susceptibilities and will be denoted as density-matrix susceptibilities α and $\tilde{\alpha}$. For imaginary frequencies (the only ones which are of interest in the context of this work), these two new quantities are defined through transition density matrices and excitation energies in the following way:

$$\alpha_{A}(1|1';2|i\omega) = 2 \sum_{K=1}^{\infty} \rho_{A}^{0 \to K}(1|1')\rho_{A}^{K \leftarrow 0}(2|2)\frac{\Delta E_{A}^{K}}{(\Delta E_{A}^{K})^{2} + \omega^{2}} \tag{14}$$

$$\tilde{\alpha}_{A}(12|1'2';3|i\omega) = 2 \sum_{K=1}^{\infty} \Gamma_{A}^{0 \to K}(12|1'2')\rho_{A}^{K \leftarrow 0}(3|3)\frac{\Delta E_{A}^{K}}{(\Delta E_{A}^{K})^{2} + \omega^{2}} \tag{15}$$

(For the monomer B, the definitions are completely analogous).

It is noteworthy that the density-matrix susceptibility α can be defined in a way that is independent of the sum-over-state expansion (eq 14). To this end, a quantity related to the polarization propagator, which will be denoted as a "half-propagator" $\langle\langle X; Y \rangle\rangle_{i\omega}^{+}$ should be defined for operators X and Y (the Y operator should be Hermitian):

$$\langle\langle X;Y \rangle\rangle_{i\omega}^{+} = -2\langle\Psi_{0}|X Re\frac{Q}{H - E_{0} - i\omega}Y\Psi_{0}\rangle \tag{16}$$

where $Q = 1 - |\Psi_{0}\rangle\langle\Psi_{0}|$ is the projection on the space orthogonal to Ψ_{0}. Then the density-matrix susceptibility can be defined as

$$\alpha(1|1';2|i\omega) = -\langle\langle \hat{\rho}(1|1');\hat{\rho}(2) \rangle\rangle_{i\omega}^{+} \tag{17}$$

where $\hat{\rho}(1)$ and $\hat{\rho}(1|1')$ are the electron density and density-matrix operators. In the algebraic approximation these operators are given by the equations:

$$\begin{aligned}
\hat{\rho}(1) &= \sum_{pq} \phi_{p}^{*}(1)\phi_{q}(1)\hat{e}_{q}^{p} \\
\hat{\rho}(1|1') &= \sum_{pq} \phi_{p}^{*}(1)\phi_{q}(1')\hat{e}_{q}^{p}
\end{aligned} \tag{18}$$

The operator $\hat{e}_{q}^{p} = a^{p}a_{q}$ entering eq 18 denotes the usual spinorbital replacement operator[48] (a^{p} and a_{p} are the creation and annihilation operators for the spinorbital ϕ_{p}). In the following, real spinorbitals will be assumed. It can be noted that a usual density susceptibility $\alpha(1,2|i\omega)$ is obtained from eq 17 if $1 = 1'$ is set.

After making use of definitions from eqs 14 and 15, the second-order exchange-dispersion energy can be expressed as

$$\begin{aligned}
E_{\text{exch-disp}}^{(2)} =\ & \frac{1}{2\pi} \int_{0}^{\infty} \int \alpha_{A}(1|1';3|i\omega)\alpha_{B}(1'|1;3'|i\omega)v(1, 1')\frac{1}{r_{33'}}d\tau_{1}d\tau_{1}'d\tau_{3}d\tau_{3}'d\omega \\
+\ & \frac{1}{2\pi} \int_{0}^{\infty} \int \alpha_{A}(1|2';3|i\omega)\tilde{\alpha}_{B}(1'2'|1'1;3'|i\omega)v(1, 1')\frac{1}{r_{33'}}d\tau_{1}d\tau_{1}'d\tau_{2}'d\tau_{3}d\tau_{3}'d\omega \\
+\ & \frac{1}{2\pi} \int_{0}^{\infty} \int \tilde{\alpha}_{A}(12|11';3|i\omega)\alpha_{B}(1'|2;3'|i\omega)v(1, 1')\frac{1}{r_{33'}}d\tau_{1}d\tau_{1}'d\tau_{2}d\tau_{3}d\tau_{3}'d\omega \\
+\ & \frac{1}{2\pi} \int_{0}^{\infty} \int \tilde{\alpha}_{A}(12|12';3|i\omega)\tilde{\alpha}_{B}(1'2'|1'2;3'|i\omega)v(1, 1')\frac{1}{r_{33'}}d\tau_{1}d\tau_{1}'d\tau_{2}d\tau_{2}'d\tau_{3}d\tau_{3}'d\omega \\
-\ & E_{\text{elst}}^{(1)}\frac{1}{2\pi} \int_{0}^{\infty} \int \alpha_{A}(2|2';3|i\omega)\alpha_{B}(2'|2;3'|i\omega)\frac{1}{r_{33'}}d\tau_{2}d\tau_{2}'d\tau_{3}d\tau_{3}'d\omega \\
+\ & \bar{P}\frac{1}{2\pi} \int_{0}^{\infty} \int \alpha_{A}(2|2;3|i\omega)\alpha_{B}(2'|2';3'|i\omega)v(2, 2')\frac{1}{r_{33'}}d\tau_{2}d\tau_{2}'d\tau_{3}d\tau_{3}'d\omega
\end{aligned} \tag{19}$$

Equation 19 describes the exchange-dispersion energy in terms of monomer properties, analogously to the Longuet-Higgins formula for the dispersion energy.

2.2. Explicitly Connected Formula for the Exchange-Dispersion Energy. Equation 19 contains two explicitly disconnected terms (the last two terms), and it is not clear from the form of this expression if they cancel with some parts of the main term (they *should* cancel since the exchange-dispersion energy is size-extensive). In order to perform the explicit cancellation, the concept of the transition density cumulant is utilized. The cumulant of the transition density matrix has been introduced in ref 42 where it has served to remove explicitly disconnected parts in the formula for the second-order exchange-induction energy. Two-electron transition density matrices $\Gamma^{0\rightarrow K}$ can be divided into five additive parts: four products of one-electron transition and ground-state density matrices and a size-extensive cumulant part:

$$\Gamma^{0\rightarrow K}(12|1'2') = \rho^{0\rightarrow K}(1|1')\rho(2|2') - \rho^{0\rightarrow K}(1|2')\rho(2|1') + \rho(1|1')\rho^{0\rightarrow K}(2|2') - \rho(1|2')\rho^{0\rightarrow K}(2|1') + \Lambda^{0\rightarrow K}(12|1'2') \quad (20)$$

The specific formulas for quantities given in eq 20 can be obtained in the algebraic approximation if the matrices $\Gamma^{0\rightarrow K}$ and $\rho^{0\rightarrow K}$ are expanded according to eqs 19 and 20 of ref 42. From eq 20 it follows that the $\tilde{\alpha}$ density-matrix susceptibility can be also decomposed into the same number of components. Four of these terms are constructed from products of the α density-matrix susceptibility and one-electron density matrices, while the fifth term is a true two-electron part of $\tilde{\alpha}$, so it will be named a *cumulant susceptibility* and denoted as λ:

$$\tilde{\alpha}(12|1'2';3|i\omega) = \alpha(1|1';3|i\omega)\rho(2|2') - \alpha(1|2';3|i\omega)\rho(2|1') + \alpha(2|2';3|i\omega)\rho(1|1') - \alpha(2|1';3|i\omega)\rho(1|2') + \lambda(12|1'2';3|i\omega) \quad (21)$$

[Note parenthetically that if the summation in eq 15 is performed over Hartree−Fock determinants, then it is restricted to singly excited configurations, so that the quantity λ in this case is equal to zero.] The partition given in eq 21 can be utilized to divide $E^{(2)}_{\text{exch-disp}}$ into the part containing only α's and ρ's (this part will be denoted as "(n)"), and the rest, having at least one cumulant-like quantity λ (denoted as "(c)"). Two terms resulting from the fourth integral of eq 19 cancel with the explicitly disconnected terms, analogously as in the approach used in ref 42 for the exchange-induction energy. The remaining expression for the exchange-dispersion energy does not contain products of integrals and is, therefore, explicitly size-extensive, provided that proper (i.e., size-extensive) density-matrix susceptibilities are utilized. For a brevity of the next formula for the exchange-dispersion energy, the following notation will be introduced: if a product $f(i|j)g(k|l)$ of functions f and g depending on electron coordinates i,j,k,l is given, then the symbol π^*_{ik} permutes the coordinates of f and g placed just before the vertical bar, while π_{jl} permutes the electron coordinates just after the vertical bar. Additionally, the symbols $\mathscr{A}^*_{ik} = 1 - \pi^*_{ik}$ and $\mathscr{A}_{jl} = 1 - \pi_{jl}$ will be defined. Then, e.g., $\pi_{12}\alpha(1|1';3|i\omega)\rho(2|2) = \alpha(1|2;3|i\omega)\rho(2|1)$, while $\pi^*_{12}\alpha(1|1;3|i\omega)\rho(2|2) = \alpha(2|1;3|i\omega)\rho(1|2)$. Finally, let π^X_{ij} and $\mathscr{A}^X_{ij}$ (with or without an asterisk) denote symbols acting on electron coordinates of α_X and ρ_X only (X = A, B). These symbols will generate the parts of the formula for the exchange-dispersion energy, which have not been written explicitly for the sake of conciseness. With these definitions, the explicitly connected formulas for the noncumulant part of the exchange-dispersion energy can be written in the following compact form:

$$\begin{aligned}
E^{(2)}_{\text{exch-disp}}(n) = & \frac{1}{2\pi}\int_0^\infty \int \alpha_A(1|1';3|i\omega)\alpha_B(1'|1;3'|i\omega)v(1,1')\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau_3 d\tau'_3 d\omega \\
& + \frac{1}{2\pi}\int_0^\infty \int v(1,1')\alpha_A(1|2';3|i\omega)\mathscr{A}^{*B}_{1'2'}\mathscr{A}^B_{1'1}\alpha_B(1'|1';3'|i\omega)\rho_B(2'|1)\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau'_2 d\tau_3 d\tau'_3 d\omega \\
& + \frac{1}{2\pi}\int_0^\infty \int v(1,1')\mathscr{A}^{*A}_{12}\mathscr{A}^A_{11'}\alpha_A(1|1;3|i\omega)\rho_A(2|1')\alpha_B(1'|2;3'|i\omega)\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau_2 d\tau_3 d\tau'_3 d\omega \\
& + \frac{1}{2\pi}\int_0^\infty \int v(1,1')(\mathscr{A}^{*A}_{12}\mathscr{A}^A_{12'}\mathscr{A}^{*B}_{1'2'}\mathscr{A}^B_{1'2} - 1 - \pi^{*A}_{12}\pi^A_{12'}, \pi^{*B}_{1'2'}, \pi^B_{1'2})\alpha_A(1|1;3|i\omega)\rho_A(2|2') \\
& \times \alpha_B(1'|1';3'|i\omega)\rho_B(2'|2)\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau_2 d\tau'_2 d\tau_3 d\tau'_3 d\omega
\end{aligned} \quad (22)$$

Similarly, the cumulant-containing part is given by the following expression:

$$\begin{aligned}
E^{(2)}_{\text{exch-disp}}(c) = & \frac{1}{2\pi}\int_0^\infty \int \alpha_A(1|2';3|i\omega)\lambda_B(1'2'|1'1;3'|i\omega)v(1,1')\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau'_2 d\tau_3 d\tau'_3 d\omega \\
& + \frac{1}{2\pi}\int_0^\infty \int \lambda_A(12|12';3|i\omega)\alpha_B(1'|2;3'|i\omega)v(1,1')\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau_2 d\tau_3 d\tau'_3 d\omega \\
& + \frac{1}{2\pi}\int_0^\infty \int v(1,1')(\mathscr{A}^{*A}_{12}\mathscr{A}^A_{12'}\alpha_A(1|1;3|i\omega)\rho_A(2|2')\lambda_B(1'2'|1'2;3'|i\omega) \\
& + \lambda_A(12|11';3|i\omega)\mathscr{A}^{*B}_{1'2'}\mathscr{A}^B_{1'2}\alpha_B(1'|1';3'|i\omega)\rho_B(2'|2) \\
& + \lambda_A(12|11';3|i\omega)\lambda_B(1'2'|1'2;3'|i\omega))\frac{1}{r_{33'}}d\tau_1 d\tau'_1 d\tau_2 d\tau'_2 d\tau_3 d\tau'_3 d\omega
\end{aligned} \quad (23)$$

Equations 22 and 23 are the main formulas of the first part of this paper.

It can be noted that a similar procedure can be employed to express the exchange-induction energy through static density-matrix susceptibilities of monomers. The corresponding formula is given in the Appendix Section.

2.3. Spinorbital Form of $E^{(2)}_{\text{exch}-\text{disp}}$. For practical applications, the formulas presented in eqs 22 and 23 should be rewritten in the algebraic approximation. If a spinorbital basis $\{\phi_p\}$ is used, then the expansions of one-electron density matrices and transition density matrices for the monomer A take the form:

$$\rho_A(1|2) = \sum_{pq} (\rho_A)^p_q \phi^*_p(1)\phi_q(2) \tag{24}$$

$$\rho^{0\rightarrow K}_A(1|2) = \sum_{pq} (\rho^{0\rightarrow K}_A)^p_q \phi^*_p(1)\phi_q(2) \tag{25}$$

where the expansion coefficients can be found from the following expressions:

$$\begin{aligned} (\rho_A)^p_q &= \langle \Psi^0_A | \hat{e}^p_q \Psi^0_A \rangle \\ (\rho^{0\rightarrow K}_A)^p_q &= \langle \Psi^0_A | \hat{e}^p_q \Psi^K_A \rangle \end{aligned} \tag{26}$$

The expressions for the monomer B are obtained by replacing the index A by B.

An expansion of the density-matrix susceptibility in a spinorbital basis leads to the following expression (we skip the index A):

$$\alpha(1|2;3|i\omega) = \sum_{p_1q_1} \sum_{p_2q_2} \Pi^{p_1p_2}_{q_1q_2}(i\omega)\phi^*_{p_1}(1)\phi_{q_1}(2)\phi^*_{p_2}(3)\phi_{q_2}(3) \tag{27}$$

where coefficients $\Pi^{p_1p_2}_{q_1q_2}(i\omega)$ are defined as

$$\Pi^{p_1p_2}_{q_1q_2}(i\omega) = 2\sum_{K=1} (\rho^{0\rightarrow K})^{p_1}_{q_1}(\rho^{K\rightarrow 0})^{p_2}_{q_2}\frac{\Delta E^K}{(\Delta E^K)^2 + \omega^2} \tag{28}$$

If real spinorbitals are considered, the remaining permutational symmetry of spinorbital indices p_2 and q_2 in eq 27 allows to rewrite the coefficients in the following form:

$$\Pi^{p_1p_2}_{q_1q_2}(i\omega) = -\langle\langle \hat{e}^{p_1}_{q_1}; \hat{g}^{p_2}_{q_2}\rangle\rangle^+_{i\omega} \tag{29}$$

where $\hat{g}^{p_2}_{q_2} = (1/2)(\hat{e}^{p_2}_{q_2} + \hat{e}^{q_2}_{p_2})$. In practice, these coefficients can be obtained by a slight modification of the algorithm for the calculation of the polarization propagators.

Six-index expansion coefficients for the cumulant susceptibilities can be defined analogously to eq 28:

$$\Lambda^{p_1p_2,p_3}_{q_1q_2,q_3}(i\omega) = 2\sum_{K=1} (\Lambda^{0\rightarrow K})^{p_1p_2}_{q_1q_2}(\rho^{K\rightarrow 0})^{p_3}_{q_3}\frac{\Delta E^K}{(\Delta E^K)^2 + \omega^2} \tag{30}$$

(for a definition of the $(\Lambda^{0\rightarrow K})^{p_1p_2}_{q_1q_2}$ expansion coefficients in eq 30 see ref 42). To proceed further let us introduce the following permutation operator:

$$P_{p_1p_2,q_1q_2} = (1 - P_{p_1p_2})(1 - P_{q_1q_2}) \tag{31}$$

where P_{rs} permutes spinorbital indices r and s, and a "chemical" notation for generalized two-electron integrals is used

$$(pq|v|rs) = \int \phi^*_p(1)\phi_q(1)v(1,2)\phi^*_r(2)\phi_s(2)\mathrm{d}\tau_1\mathrm{d}\tau_2 \tag{32}$$

A pure two-electron repulsion integral is denoted as

$$(pq|rs) = \int \phi^*_p(1)\phi_q(1)\frac{1}{r_{12}}\phi^*_r(2)\phi_s(2)\mathrm{d}\tau_1\mathrm{d}\tau_2 \tag{33}$$

Combining eqs 24, 27, 32, and 33, one can transform the noncumulant part of the exchange-dispersion energy to the following spinorbital form:

$$
\begin{aligned}
E^{(2)}_{\text{exch}-\text{disp}}(n) &= \frac{1}{2\pi}\Big\{ \int_0^\infty {}^A\Pi^{p_1p_3}_{q_1q_3}(i\omega){}^B\Pi^{r_1r_3}_{s_1s_3}(i\omega)\mathrm{d}\omega(p_1s_1|v|r_1q_1) \\
&+ (p_1s_2|v|r_1s_1)S^{q_1}_{r_2}P_{r_1r_2,s_1s_2}\int_0^\infty {}^A\Pi^{p_1p_3}_{q_1q_3}(i\omega){}^B\Pi^{r_2r_3}_{s_2s_3}(i\omega)\mathrm{d}\omega(\rho_B)^{r_1}_{s_1} \\
&+ (p_1q_1|v|r_1q_2)S^{s_1}_{p_2}P_{p_1p_2,q_1q_2}\int_0^\infty {}^A\Pi^{p_2p_3}_{q_2q_3}(i\omega){}^B\Pi^{r_1r_3}_{s_1s_3}(i\omega)\mathrm{d}\omega(\rho_A)^{p_1}_{q_1} \\
&+ (p_1q_1|v|r_1s_1)S^{s_2}_{p_2}S^{q_2}_{r_2}(P_{p_1p_2,q_1q_2}P_{r_1r_2,s_1s_2} - 1 - P_{p_1p_2}P_{q_1q_2}P_{r_1r_2}P_{s_1s_2}) \\
&\times \int_0^\infty {}^A\Pi^{p_2p_3}_{q_2q_3}(i\omega){}^B\Pi^{r_2r_3}_{s_2s_3}(i\omega)\mathrm{d}\omega(\rho_A)^{p_1}_{q_1}(\rho_B)^{r_1}_{s_1}\Big\}(p_3q_3|r_3s_3)
\end{aligned}
\tag{34}
$$

An analogous derivation for the cumulant part leads to the expression:

$$
\begin{aligned}
E^{(2)}_{\text{exch}-\text{disp}}(c) &= \frac{1}{2\pi}\Big\{ \int_0^\infty {}^A\Pi^{p_1p_3}_{q_1q_3}(i\omega){}^B\Lambda^{r_1r_2,r_3}_{s_1s_2,s_3}(i\omega)\mathrm{d}\omega(p_1s_2|v|r_1s_1)S^{q_1}_{r_2} \\
&+ \int_0^\infty {}^A\Lambda^{p_1p_2,p_3}_{q_1q_2,q_3}(i\omega){}^B\Pi^{r_1r_3}_{s_1s_3}(i\omega)\mathrm{d}\omega(p_1q_1|v|r_1q_2)S^{s_1}_{p_2} \\
&+ (p_1q_1|v|r_1s_1)S^{s_2}_{p_2}S^{q_2}_{r_2}\Big[P_{p_1p_2,q_1q_2}\int_0^\infty {}^A\Pi^{p_2p_3}_{q_2q_3}(i\omega){}^B\Lambda^{r_1r_2,r_3}_{s_1s_2,s_3}(i\omega)\mathrm{d}\omega(\rho_A)^{p_1}_{q_1} \\
&+ P_{r_1r_2,s_1s_2}\int_0^\infty {}^A\Lambda^{p_1p_2,p_3}_{q_1q_2,q_3}(i\omega){}^B\Pi^{r_2r_3}_{s_2s_3}(i\omega)\mathrm{d}\omega(\rho_B)^{r_1}_{s_1} \\
&+ \int_0^\infty {}^A\Lambda^{p_1p_2,p_3}_{q_1q_2,q_3}(i\omega){}^B\Lambda^{r_1r_2,r_3}_{s_1s_2,s_3}(i\omega)\mathrm{d}\omega\Big]\Big\}(p_3q_3|r_3s_3)
\end{aligned}
\tag{35}
$$

In eqs 34 and 35 the summation over repeated indices is assumed (Einstein convention).

2.4. CC Exchange-Dispersion energy. The presented formalism is so far completely general, and any well-defined method for polarization propagators can be used to produce density-matrix susceptibilities for eqs 34 and 35. In particular, the energy $E^{(200)}_{\text{exch}-\text{disp}} \equiv E^{(2)}_{\text{exch}-\text{disp}}(\text{UCHF})$ can be obtained if the uncoupled HF (UCHF) propagators are used in eq 35. When the electron-correlated polarization propagators are utilized instead, the exchange-dispersion energy with the inclusion of the intramonomer correlation effect is obtained. In this work, the intramonomer correlation is studied on the CC theory level. The polarization propagators used here are obtained from time-independent CC theory developed by Moszynski et al.[49] The general equation for the imaginary-frequency CC polarization propagator from this paper adapted for our purposes takes the form:

$$
\langle\langle Y;X\rangle\rangle^+_{i\omega} = 2\langle e^{-S}e^{T^\dagger}Ye^{-T^\dagger}e^S\Phi|\hat{\mathscr{P}}(e^{S^\dagger}Re\Omega^X(i\omega)e^{-S^\dagger})\Phi\rangle
\tag{36}
$$

where Φ denotes the reference determinant, $\hat{\mathscr{P}}$ stands for the superoperator $\hat{\mathscr{P}}_1 + \hat{\mathscr{P}}_2 + \hat{\mathscr{P}}_3 + ...$, defined in such a way that $\hat{\mathscr{P}}_n(Z)$ projects on the n-tuple excitation part of an operator Z,[50] T is the usual amplitude operator of CC theory, and S is the excitation operator, $S = S_1 + S_2 + S_3 + ...$, introduced in refs 9 and 50 through the formula:

$$
e^S\Phi = \frac{1}{\langle e^T\Phi|e^T\Phi\rangle}e^{T^\dagger}e^T\Phi
\tag{37}
$$

Finally, $\Omega^X(i\omega)$ stands for the first-order perturbed operator,[49] depending linearly on the perturbation X. The time-independent theory of the CC polarization propagator has been recently implemented and studied on the CCSD level.[51] Several variants of this method have been examined in ref 51 and finally the CCSD(3) model has been chosen as optimal for many-electron cases. [The CCSD(n) model includes the minimum number of terms from the expanded form of eqs 36 and 37, which is necessary to include all CCSD terms contributing on at least the $\mathcal{O}(W^n)$ order.] The CCSD(n) propagators have been already used for the induction, exchange-induction[42] and exchange-dispersion energies[39] and have been shown to produce very accurate results in comparison to the benchmark values for the helium dimer.[52] In the following, the number in parentheses will be dropped if CCSD(3) propagators are used for SAPT corrections. The experience gained from the investigation of the cumulant parts of the first-order exchange and second-order exchange-induction energies tells us that the cumulant part constitutes usually only a few percent of the total correction.[41,42] Since the exchange-dispersion energy, unlike the exchange-induction, is small in comparison to its polarization counterpart (it usually amounts to only 5−15% of the dispersion energy in the van der Waals minimum); the cumulant part of the exchange-dispersion energy will be not implemented in the present study.

A closer examination of eq 35 reveals that the calculation of $E^{(2)}_{\text{exch}-\text{disp}}(n)$ scales with the sixth power of a dimension of the orbital basis N_{AO}, while in eq 35 the computational cost of the first and second terms scales with the seventh and the last term scales with the eighth power of N_{AO}. The high scaling of eq 35 is an additional argument to skip this part of the exchange-dispersion energy in practical implementations (even if programmed, it could not be used but for the smallest monomers anyway). Note parenthetically that the present formulation of the exchange-dispersion energy allows monomers to be described in different basis sets (the so-called monomer-centered basis set approach[53]).

2.5. Density Fitting of Density-Matrix Susceptibilities. The cost of a calculation of full four-index CCSD propagators scales with the eighth power of the molecular size,[39] since: (i) one has to calculate $(N_{\text{AO}} + 1)N_{\text{AO}}/2$ responses to the perturbations being the $\hat{G}^p_q = 1/2(\hat{E}^p_q + \hat{E}^q_p)$ operators, where $\hat{E}^p_q = \hat{e}^{p\alpha}_{q\alpha} + \hat{e}^{p\beta}_{q\beta}$ are the usual orbital replacement operators[48,54] (in this chapter

small letters will denote orbital indices); and (ii) the cost of one response calculation scales as iterative $\mathcal{O}(v^4 o^2)$, where v and o are dimensions of virtual and occupied orbital subspaces, respectively.

In ref 39 Korona and Jeziorski have shown that by performing a fitting of CCSD density susceptibilities to some suitably selected auxiliary basis $\{\chi_K\}$ of a dimension N_{aux} (usually N_{aux} is only 2−3 times larger than N_{AO}), the computational cost for the CC dispersion energy can be reduced by one order of magnitude. To this end, the density susceptibility α has been expanded as

$$\alpha(\mathbf{r}_1, \mathbf{r}_2|\omega) = -\sum_{KL} \langle\langle \hat{\chi}_K; \hat{\chi}_L \rangle\rangle_\omega \chi_K(\mathbf{r}_1)\chi_L(\mathbf{r}_2) \tag{38}$$

where the operators $\hat{\chi}_K$ are defined as

$$\hat{\chi}_K = \sum_{pq} D_{pq}^K \hat{E}_q^p \tag{39}$$

with coefficients D_{pq}^K obtained from a fitting of orbital products with the auxiliary basis:

$$\phi_p(\mathbf{r})\phi_q(\mathbf{r}) = \sum_K D_{pq}^K \chi_K(\mathbf{r}) \tag{40}$$

In the following, capital letters will be reserved for indices of the auxiliary basis set.

For the density-matrix susceptibilities it is still possible to perform a density fitting procedure on the orbital product $\phi_{p_2}(\mathbf{r}_3)\phi_{q_2}(\mathbf{r}_3)$ from eq 27 (a summation over spin is assumed to have been already performed), which leads to the following asymmetric form of α:

$$\alpha(\mathbf{r}_1|\mathbf{r}_2;\mathbf{r}_3|i\omega) = -\sum_{pq}\sum_K \langle\langle \hat{E}_q^p; \hat{\chi}_K \rangle\rangle_{i\omega}^+ \phi_p(\mathbf{r}_1)\phi_q(\mathbf{r}_2)\chi_K(\mathbf{r}_3) \tag{41}$$

The expansion coefficients $\langle\langle \hat{E}_q^p; \hat{\chi}_K \rangle\rangle_{i\omega}^+$ will be denoted as $\Pi_{qK}^p(i\omega)$. Additionally, in the next formulas a letter A and B over Π will differentiate between propagators of the monomer A and B, respectively.

After performing the summation over spins in eq 35 and making use of eq 41, the following formula is obtained for the exchange-dispersion energy with the DF approximation:

$$
\begin{aligned}
E_{\text{exch-disp}}^{(2)}(\text{n}) =\ & \frac{1}{\pi}\int_0^\infty \overset{A}{\Pi}_{q_1 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_1}(i\omega)d\omega(N|M)(p_1 s_1|v|r_1 q_1) \\
+\ & \frac{1}{2\pi}\int_0^\infty \overset{A}{\Pi}_{q_1 N}^{p_1}(i\omega)[2\overset{B}{\Pi}_{s_2 M}^{r_2}(i\omega)d\omega(\rho_B)_{s_1}^{r_1} + 2\overset{B}{\Pi}_{s_1 M}^{r_1}(i\omega)d\omega(\rho_B)_{s_2}^{r_2} \\
-\ & \overset{B}{\Pi}_{s_1 M}^{r_2}(i\omega)d\omega(\rho_B)_{s_2}^{r_1} - \overset{B}{\Pi}_{s_2 M}^{r_1}(i\omega)d\omega(\rho_B)_{s_1}^{r_2}](N|M)(p_1 s_2|v|r_1 s_1)S_{r_2}^{q_1} \\
+\ & \frac{1}{2\pi}\int_0^\infty [2\overset{A}{\Pi}_{q_2 N}^{p_2}(i\omega)(\rho_A)_{q_1}^{p_1} + 2\overset{A}{\Pi}_{q_1 N}^{p_1}(i\omega)(\rho_A)_{q_2}^{p_2} - \overset{A}{\Pi}_{q_1 N}^{p_2}(i\omega)(\rho_A)_{q_2}^{p_1} \\
-\ & \overset{A}{\Pi}_{q_2 N}^{p_1}(i\omega)(\rho_A)_{q_1}^{p_2}]\overset{B}{\Pi}_{s_1 M}^{r_1}(i\omega)d\omega(N|M)(p_1 q_1|v|r_1 q_2)S_{p_2}^{s_1} \\
+\ & \frac{1}{4\pi}\int_0^\infty \Big[4\overset{A}{\Pi}_{q_2 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_1}(i\omega)(\rho_A)_{q_1}^{p_1}(\rho_B)_{s_2}^{r_2} + 4\overset{A}{\Pi}_{q_1 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_2}(i\omega)(\rho_A)_{q_2}^{p_2}(\rho_B)_{s_1}^{r_1} \\
-\ & 2\overset{A}{\Pi}_{q_2 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_2}(i\omega)(\rho_A)_{q_1}^{p_1}(\rho_B)_{s_2}^{r_1} - 2\overset{A}{\Pi}_{q_2 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_1}(i\omega)(\rho_A)_{q_1}^{p_1}(\rho_B)_{s_1}^{r_2} \\
-\ & 2\overset{A}{\Pi}_{q_1 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_2}(i\omega)(\rho_A)_{q_2}^{p_2}(\rho_B)_{s_2}^{r_1} - 2\overset{A}{\Pi}_{q_1 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_1}(i\omega)(\rho_A)_{q_2}^{p_2}(\rho_B)_{s_1}^{r_2} \\
-\ & 2\overset{A}{\Pi}_{q_1 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_2}(i\omega)(\rho_A)_{q_2}^{p_1}(\rho_B)_{s_1}^{r_1} - 2\overset{A}{\Pi}_{q_1 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_1}(i\omega)(\rho_A)_{q_2}^{p_1}(\rho_B)_{s_2}^{r_2} \\
-\ & 2\overset{A}{\Pi}_{q_2 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_2}(i\omega)(\rho_A)_{q_1}^{p_2}(\rho_B)_{s_1}^{r_1} - 2\overset{A}{\Pi}_{q_2 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_1}(i\omega)(\rho_A)_{q_1}^{p_2}(\rho_B)_{s_2}^{r_2} \\
+\ & \overset{A}{\Pi}_{q_1 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_2}(i\omega)(\rho_A)_{q_2}^{p_1}(\rho_B)_{s_2}^{r_1} + \overset{A}{\Pi}_{q_1 N}^{p_2}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_1}(i\omega)(\rho_A)_{q_2}^{p_1}(\rho_B)_{s_1}^{r_2} \\
\times\ & \overset{A}{\Pi}_{q_2 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_1 M}^{r_2}(i\omega)(\rho_A)_{q_1}^{p_2}(\rho_B)_{s_2}^{r_1} + \overset{A}{\Pi}_{q_2 N}^{p_1}(i\omega)\overset{B}{\Pi}_{s_2 M}^{r_1}(i\omega)(\rho_A)_{q_1}^{p_2}(\rho_B)_{s_1}^{r_2}\Big]d\omega(N|M)(p_1 q_1|v|r_1 s_1)S_{p_2}^{s_2}S_{r_2}^{q_2}
\end{aligned}
\tag{42}
$$

where $(N|M) = \int \chi_N(\mathbf{r}_1)r_{12}^{-1}\chi_M(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2$, and the summation over spins have been performed as in ref 40. Note that the dispersion energy can also be expressed through mixed three-index propagators:

$$E_{\text{disp}}^{(2)} = -\frac{2}{\pi}\int_0^\infty \overset{A}{\Pi}_{qN}^p(i\omega)\overset{B}{\Pi}_{sM}^r(i\omega)d\omega(N|M)(pq|rs) \tag{43}$$

It should be stressed that this equation is not bitwise compatible with the formula for the dispersion energy obtained with fully density-fitted density susceptibilities (see eq 38), but the differences should be the same order, as the accuracy of density fitting for the $(pq|rs)$ integrals (they would dissapear if DF were exact). The non-DF orbital equations can be retrieved from

Exchange-Dispersion Energy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2671**

eq 42 by replacing auxiliary indices $N(M)$ by pairs of indices $p_3q_3(r_3s_3)$. The integral over ω is calculated numerically (for details see ref 51). In order to obtain a three-index matrix $\Pi_{qN}^p(i\omega)$, responses to N_{aux} perturbation operators $\hat{\chi}_K$ are calculated first. With $\Omega^{\hat{\chi}_K}(i\omega)$ responses in hand, a square matrix of coefficients with indices p,q running from 1 to N_{AO} can be easily obtained from eq 15 of ref 39 by inserting $\hat{E}_q^p$ as the second perturbing operator. The computational cost of one $\Omega^{\hat{\chi}_K}(i\omega)$ operator for the CCSD case scales as $\mathcal{O}(o^2v^4)$, so for N_{aux} such operators the scaling with seventh power of the molecular size is obtained.

Note that the calculation of the exchange-dispersion energy from ready propagators is also accelerated by one order of the orbital basis size if DF is used: the calculation scales as $\mathcal{O}(N_{\text{AO}}^6)$ for the no-DF case and as $\mathcal{O}(N_{\text{AO}}^4 N_{\text{aux}})$ if DF is applied. However, this fact is of limited importance, since the calculation of the CCSD propagators is the most expensive part of the program, anyway.

The one-electron density matrices used in this paper have been obtained from CC theory for the expectation value developed by Jeziorski and Moszynski[50] and implemented on the CCSD level by Korona and Jeziorski.[37] The exchange-dispersion energy obtained from eq 35 with the CCSD(n) propagators[51] and with XCCSD-3 density matrices[37] will be denoted simply as $E_{\text{exch-disp}}^{(2)}(\text{CCSD})$.

2.6. Implementation. Both DF and non-DF formulas for the noncumulant part of the exchange-dispersion energy have been implemented as a local extension of the MOLPRO code.[55] The two- and three-index Coulomb integrals, coded in MOLPRO for purposes of density fitted MP2[56] and local CCSD[57] methods, were used to obtain the $\hat{\chi}_K$ operators and to calculate the exchange-dispersion and dispersion energies from eqs 42 and 43. Since the summation over orbital indices in the CCSD half-propagator coefficients is not limited, there is no gain in performing the transformation to the molecular orbital basis in this case. Therefore, the formulas have been implemented in atomic orbitals (AO). Noteworthy, such an implementation makes it easy to extend the program to monomer-centered basis sets in the future. As already mentioned, the new code yields the $E_{\text{exch-disp}}^{(2)}(\text{UCHF})$ energy if the UCHF polarization half-propagators are inserted into eq 35. Analogously, by using the coupled HF (CHF) half-propagators, the $E_{\text{exch-disp}}^{(2)}(\text{CHF})$ energy is obtained. The calculations for the latter correction has not appeared in the literature yet, however, it has been derived and programmed in the MOLPRO code as a byproduct of SAPT(DFT) by Hesselmann et al.[27,29] The authors of these references focused however on the coupled perturbed Kohn–Sham (CKS) dispersion and exchange-dispersion energies and did not present any calculations for the CHF case. It should be stressed that in the SAPT(DFT) approach programmed in the SAPT code,[13,31,32] the uncoupled KS (UCKS) method is used for the exchange-induction and exchange-dispersion energies, and afterward scaling formulas are applied to estimate the coupled part of these corrections, e.g., for the exchange-dispersion energy:

$$\tilde{E}_{\text{exch-disp}}^{(2)}(\text{CKS}) = \frac{E_{\text{disp}}^{(2)}(\text{CKS})}{E_{\text{disp}}^{(2)}(\text{UCKS})} E_{\text{exch-disp}}^{(2)}(\text{UCKS}) \tag{44}$$

Since the coupled variants of the interaction energy components are usually more accurate that the uncoupled ones, the SAPT(DFT) approach presented by Hesselmann et al.[27,29] should be preferred over the approach based on the UCKS method.[31,32]

Since the new code has been programmed in the AO basis, it has been easy to examine its correctness by inserting the AO-transformed UCHF (or CHF) quantities (density-matrix susceptibilities and one-electron density matrices) instead of the CCSD ones and by checking if the obtained value is equal to the $E_{\text{exch-disp}}^{(2)}(\text{UCHF})$ (or $E_{\text{exch-disp}}^{(2)}(\text{CHF})$) energy obtained from the SAPT(DFT) program of Hesselmann and Jansen.

For an efficient implementation of the exchange-dispersion energy, it is useful to define the following intermediate:

$$X_{qs}^{pr} = \frac{1}{2\pi} \int_0^\infty \overset{\text{A}}{\Pi}_{qN}^p(i\omega) \overset{\text{B}}{\Pi}_{sM}^r(i\omega) \mathrm{d}\omega (N|M) \tag{45}$$

This quantity for fixed two indices is calculated on the fly and then used in the calculation of the corresponding part of the exchange-dispersion energy. It is possible to switch on the paging over frequencies if all three-index objects do not fit into memory. With the X_{qs}^{pr} intermediate defined, the formula given in eq 42 can be rewritten as

$$
\begin{aligned}
E_{\text{exch-disp}}^{(2)}(n) = {} & 2X_{q_1s_1}^{p_1r_1}(p_1s_1|v|r_1q_1) \\
& + [2X_{q_1s_2}^{p_1r_2}(\rho_\text{B})_{s_1}^{r_1} + 2X_{q_1s_1}^{p_1r_1}(\rho_\text{B})_{s_2}^{r_2} - X_{q_1s_1}^{p_1r_2}(\rho_\text{B})_{s_2}^{r_1} - X_{q_1s_2}^{p_1r_1}(\rho_\text{B})_{s_1}^{r_2}](p_1s_2|v|r_1s_1)S_{r_2}^{q_1} \\
& + [2X_{q_2s_1}^{p_2r_1}(\rho_\text{A})_{q_1}^{p_1} + 2X_{q_1s_1}^{p_1r_1}(\rho_\text{A})_{q_2}^{p_2} - X_{q_1s_1}^{p_2r_1}(\rho_\text{A})_{q_2}^{p_1} - X_{q_2s_1}^{p_1r_1}(\rho_\text{A})_{q_1}^{p_2}](p_1q_1|v|r_1q_2)S_{p_2}^{s_1} \\
& + \frac{1}{2}[4X_{q_2s_1}^{p_2r_1}(\rho_\text{A})_{q_1}^{p_1}(\rho_\text{B})_{s_2}^{r_2} + 4X_{q_1s_2}^{p_1r_2}(\rho_\text{A})_{q_2}^{p_2}(\rho_\text{B})_{s_1}^{r_1} \\
& - 2X_{q_2s_1}^{p_2r_2}(\rho_\text{A})_{q_1}^{p_1}(\rho_\text{B})_{s_2}^{r_1} - 2X_{q_2s_2}^{p_2r_1}(\rho_\text{A})_{q_1}^{p_1}(\rho_\text{B})_{s_1}^{r_2} - 2X_{q_1s_1}^{p_1r_2}(\rho_\text{A})_{q_2}^{p_2}(\rho_\text{B})_{s_2}^{r_1} - 2X_{q_1s_2}^{p_1r_1}(\rho_\text{A})_{q_2}^{p_2}(\rho_\text{B})_{s_1}^{r_2} \\
& - X_{q_1s_1}^{p_2r_2}(\rho_\text{A})_{q_2}^{p_1}(\rho_\text{B})_{s_2}^{r_1} - 2X_{q_1s_1}^{p_2r_1}(\rho_\text{A})_{q_2}^{p_1}(\rho_\text{B})_{s_2}^{r_2} - 2X_{q_2s_2}^{p_1r_2}(\rho_\text{A})_{q_1}^{p_2}(\rho_\text{B})_{s_1}^{r_1} - 2X_{q_2s_1}^{p_1r_1}(\rho_\text{A})_{q_1}^{p_2}(\rho_\text{B})_{s_2}^{r_2} \\
& + X_{q_1s_1}^{p_2r_2}(\rho_\text{A})_{q_2}^{p_1}(\rho_\text{B})_{s_2}^{r_1} + X_{q_1s_2}^{p_2r_1}(\rho_\text{A})_{q_1}^{p_1}(\rho_\text{B})_{s_1}^{r_2} \\
& + X_{q_2s_1}^{p_1r_2}(\rho_\text{A})_{q_1}^{p_2}(\rho_\text{B})_{s_2}^{r_1} + X_{q_2s_2}^{p_1r_1}(\rho_\text{A})_{q_1}^{p_2}(\rho_\text{B})_{s_1}^{r_2}](p_1q_1|v|r_1s_1)S_{p_2}^{s_2}S_{r_2}^{q_2}
\end{aligned}
\tag{46}
$$

A ready-to-program version of eq 46 is presented in the Appendix Section.

Equation 46 resembles closely the UCHF formula for the exchange-dispersion energy.[7] In particular, the intermediate X is an analog of the mixed double amplitude from ref 7. This correspondence is examined in detail for one selected term in the Appendix Section.

It should be noted that an intermediate similar to X_{rs}^{pq} appears in the implementation of DF-DFT-SAPT,[29] where it is also constructed from three-index objects (partially density-fitted CKS half-propagators). It should be stressed, however, that the present approach derives rigorously the general expression of the exchange-dispersion energy in a form of density-matrix susceptibilities and density matrices, while in ref 29 a theoretical background of a replacement of the UCHF quantities by CKS or CHF ones has not been presented.

3. Results and Discussion

3.1. Computational Details. To illustrate the performance of the method presented in the previous sections several test calculations have been performed.

At the beginning the quality of DF for the exchange-dispersion energy has been investigated for the van der Waals complex of two neon atoms. In this test the distances between atoms range from 3 to 9 bohr, thus covering repulsive, minimum, and long-range regions of PES. The augmented correlation-consistent aug-cc-pVDZ[58–60] basis has been used in this case because of the necessity of performing the calculations without the density fitting approximation (scaling with the eighth power of the molecular size). The auxiliary basis sets proposed in refs 61 and 62 were utilized. The basis sets optimized for the MP2 correlation energy will be denoted as aug-cc-pVXZ/MP2fit ($X = D,T,Q$), while other basis sets optimized for the Coulomb and exchange integrals will be denoted as cc-pVXZ/JKfit. The DF calculations for a larger aug-cc-pVTZ orbital basis have been also performed.

The new approach has been then applied to the helium dimer at various distances used previously to produce an accurate PES for this van der Waals complex.[52,63] The same basis set (DC147) as in refs 52 and 63 has been used. A large auxiliary basis (aug-cc-pVQZ/MP2fit)[61,62] augmented with the functions on the midbond (see ref 39 for details) has served for a density fitting in this case. Additionally, for the $r = 5.6$ bohr a smaller DC77 basis from ref 52 has been used to calculate the exchange-dispersion energy without density fitting. The numerical quadrature with 20 integration points has been applied for the neon and helium dimers in order to eliminate uncertainties related to the quality of the numerical integration when comparing no-DF and DF energies.

Finally, the new approach has been used to calculate the exchange-dispersion and dispersion energies for selected noncovalent complexes at geometries close to global minima. These complexes are representative to various types of dominant interactions at minimum, ranging from polar systems to dispersion-bound species. The selected complexes, along with the geometry references, are the following: $(H_2O)_2$,[7] $(HF)_2$,[7] $(N_2)_2$,[64] $(CO)_2$,[65] $CO-H_2O$,[38,66] Ar_2,[31] $He-HF$, and $Ne-Ar$. For the Ne−Ar complex, a distance of 3.65 Å between atoms has been used. For the He−HF complex a distance between the helium atom and the center of mass of the HF molecule ($R_{H-F} = 1.7328$ bohr) has been set to 6.5 bohr. Some of these complexes have been previously used to study the intramonomer correlation effects for the induction[38] and dispersion[39] energies. For all cases,

the aug-cc-pVTZ orbital basis and the corresponding aug-cc-pVTZ/MP2fit basis sets have been used. The 8-point numerical quadrature has been utilized.

The core electrons (1s for C, N, O, F, and Ne and 1s2s2p for Ar) have been frozen in the CC calculations. All results have been produced with a local development version of the MOLPRO suite of codes. The SAPT(DFT) results have been obtained by using the PBE0 functional[67,68] with the asymptotic correction of Grüning et al.[69] and utilizing the code available in MOLPRO and developed by Hesselmann and Jansen.[24–27] The ionization potentials needed for the calculation of the asymptotic corrections have been obtained from ref 70.

3.2. Discussion of the Results. Since the calculation of the CCSD density-matrix susceptibility without density fitting scales with the eighth power of the molecular size, the possibility of obtaining a good quality DF-CCSD exchange-dispersion energies is crucial for the range of the applicability of the new method. Therefore, it is important to examine errors introduced by the DF approximation. Table 2 presents the dispersion and exchange-dispersion energies for the Ne$_2$ complex calculated from CCSD(3) propagators, obtained either without DF or with DF for various auxiliary basis sets. In the table, the DF results are followed by the percent errors with respect to the exact (non-DF) value. It can be seen from Table 2 that the dispersion energy is not very sensitive to a quality of auxiliary basis sets. All basis sets (even cc-pVTZ/JKfit) produce results of a similar quality with errors on the fourth digit in a majority of cases. [Note parenthetically that the dispersion energies presented here were obtained from eq 43, so they are not bitwise compatible with values from ref 39 calculated from two-index propagators.] The exchange-dispersion energy is, however, more sensitive to the quality of auxiliary basis sets, especially for large distances. From the results presented in Table 2, it can be seen that acceptable errors of a couple percent for the whole range of distances are produced by the aug-cc-pVXZ/MP2fit, $X = T,Q$ basis sets only. In view of rather substantial errors of the DF exchange-dispersion for the largest distances considered, it is interesting to examine the total error introduced by DF for the sum of the dispersion and exchange-dispersion energies. The corresponding errors for the aug-cc-pVDZ/MP2fit basis set are presented in the last line of the results for the aug-cc-pVDZ orbital basis. It turns out that errors of the exchange-dispersion energy at the large-R region are not important since contributions from the exchange-dispersion energy in this region are negligible anyway. On the other hand, for the repulsive-wall region the error caused by density fitting applied to $E_{\text{exch-disp}}^{(2)}$

Exchange-Dispersion Energy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2673**

Table 2. Dispersion and Exchange-Dispersion Energies for Two Neon Atoms (Upper Numbers) Computed from Density-Fitted CCSD(3) Density-Matrix Susceptibilities and Their Percentage Errors[a]

basis set	distance						
	3.0	4.0	5.0	6.0	7.0	8.0	9.0
aug-cc-pVDZ							
$E_{disp}^{(2)}$							
no-DF	−11.645	−2.2700	−0.4732	−0.13178	−0.04750	−0.020144	−0.009601
aug-cc-pVDZ/MP2fit	−11.649	−2.2702	−0.4731	−0.13176	−0.04749	−0.020139	−0.009601
	−0.032	−0.0094	0.013	0.010	0.0082	0.021	0.0085
aug-cc-pVTZ/MP2fit	−11.645	−2.2699	−0.4732	−0.13165	−0.04749	−0.020141	−0.009599
	0.0065	0.0043	−0.0032	0.093	0.010	0.012	0.029
aug-cc-pVQZ/MP2fit	−11.644	−2.2698	−0.4732	−0.13178	−0.04750	−0.020147	−0.009603
	0.0077	0.0087	−0.0035	−0.0046	−0.0086	−0.017	−0.017
cc-pVTZ/JKfit	−11.638	−2.2663	−0.4732	−0.13188	−0.04752	−0.020141	−0.009596
	0.060	0.16	−0.0007	−0.080	−0.049	0.011	0.056
cc-pVQZ/JKfit	−11.640	−2.2681	−0.4733	−0.13189	−0.04752	−0.020142	−0.0095963
	0.045	0.081	−0.024	−0.088	−0.057	0.0070	0.053
$E_{exch-disp}^{(2)}$							
no-DF	4.619	0.5582	0.05271	0.004531	0.000445	0.0000448	0.00000549
aug-cc-pVDZ/MP2fit	4.610	0.5559	0.05256	0.004511	0.000436	0.0000411	0.00000452
	0.20	0.41	0.29	0.43	2.14	8.09	18.
aug-cc-pVTZ/MP2fit	4.616	0.5573	0.05265	0.004608	0.000453	0.0000458	0.00000588
	0.067	0.16	0.11	−1.71	−1.78	−2.29	−7.04
aug-cc-pVQZ/MP2fit	4.618	0.5578	0.05266	0.004525	0.000443	0.0000442	0.00000539
	0.023	0.079	0.088	0.13	0.37	1.29	1.84
cc-pVTZ/JKfit	4.604	0.5528	0.05222	0.004368	0.000387	0.0000286	0.00000194
	0.32	0.96	0.94	3.59	13.	36.	65.
cc-pVQZ/JKfit	4.612	0.5557	0.05239	0.004383	0.000389	0.0000287	0.00000194
	0.16	0.44	0.60	3.26	13.	36.	65.
percent error of $E_{disp}^{(2)} + E_{exch-disp}^{(2)}$							
aug-cc-pVDZ/MP2fit	−0.18	−0.14	−0.021	−0.0044	−0.012	0.0025	−0.0016
aug-cc-pVTZ							
$E_{disp}^{(2)}$							
aug-cc-pVDZ/MP2fit	−15.228	−2.915	−0.6669	−0.18267	−0.06388	−0.026839	−0.012734
aug-cc-pVTZ/MP2fit	−15.298	−2.924	−0.6640	−0.18290	−0.06390	−0.026838	−0.012735
aug-cc-pVQZ/MP2fit	−15.300	−2.924	−0.6642	−0.18298	−0.06392	−0.026845	−0.012739
cc-pVTZ/JKfit	−15.286	−2.918	−0.6627	−0.18279	−0.06386	−0.026820	−0.012726
$E_{exch-disp}^{(2)}$							
aug-cc-pVDZ/MP2fit	5.35	0.643	0.0707	0.00679	0.00059	0.000055	0.0000055
aug-cc-pVTZ/MP2fit	5.43	0.655	0.0721	0.00701	0.00061	0.000059	0.0000062
aug-cc-pVQZ/MP2fit	5.44	0.656	0.0724	0.00699	0.00061	0.000060	0.0000064
cc-pVTZ/JKfit	5.39	0.642	0.0699	0.00658	0.00053	0.000043	0.0000032

[a] Lower numbers, for the smaller orbital basis only, relative to the results obtained without density fitting (denoted as no-DF). Energies are in millihartrees, distances are in bohr.

determines the total error of the sum of the dispersion and exchange-dispersion energies. However, differences between the DF and no-DF results of order of a few tenths of percent are acceptable in most applications. It can be, therefore, recommended that for the aug-cc-pV*XZ* basis set just the corresponding MP2fit auxiliary basis can be safely used for the SAPT(CCSD) calculations, unless we are specifically interested in the exchange-dispersion energy (for the latter case the aug-cc-pVQZ/MP2fit auxiliary basis should be utilized).

In Table 2 the DF-CCSD dispersion and exchange-dispersion energies for a larger aug-cc-pVTZ orbital basis are also presented. The results show that the exchange-dispersion energies calculated with the aug-cc-pVDZ/MP2fit and aug-cc-pVTZ/MP2fit basis sets do not differ much from each other, therefore the smaller auxiliary basis can also be used with this orbital basis.

The DF-CCSD exchange-dispersion energy for the helium dimer is shown in Table 3. The full configuration interaction (FCI) exchange-dispersion energies within the S^2 approxima-

tion listed in this table are the unpublished results collected during the calculations performed to obtain the PES for the interaction of two helium atoms.[52,63] The CCSD(4) propagators have been utilized in this case,[51] which are practically exact (equivalent to FCI) for two-electron monomers (actually, it has been checked out that the results obtained with the CCSD(3) and CCSD(4) models differ by less than 0.1% for all values listed in the table). As discussed in ref 51, that differences between the DF-CCSD and FCI dispersion energies can be attributed exclusively to the DF error. However, for the case of the density-fitted exchange-dispersion energy two sources of errors are possible: one stemming from density fitting and the second one related to the neglect of the cumulant contribution (see eq 35). The experience gained from the investigation of the Ne_2 complex tells us that the DF errors can be of order of 1−2% percent in the minimum region. Therefore, large differences between the $E_{exch-disp}^{(2)}$(CCSD) and $E_{exch-disp}^{(2)}$(FCI) corrections for distances from 3.0 to 4.0 bohr (corresponding to a highly repulsive region of PES) should be attributed rather to the cumulant

Table 3. Second-Order Exchange-Dispersion Energies in Various Approximations for the Helium Dimer[a]

method	R								
	3.0	3.5	4.0	4.5	5.0	5.3	5.6	6.0	7.0
FCI	627.6	212.7	73.06	24.98	8.453	4.386	2.265	0.9316	0.09718
DF-CCSD(4)	479	191	70.5	25.0	8.59	4.47	2.31	0.950	0.0975
	−24	−10	−3.5	−0.14	1.6	2.0	2.1	1.9	0.4
UCHF	349.3	139.7	51.29	17.93	6.075	3.139	1.611	0.6557	0.06637
	−44	−34	−29	−28	−28	−28	−29	−30	−32
CHF	423.4	171.2	63.59	22.50	7.710	4.008	2.068	0.8476	0.08691
	−33	−20	−13	−10	−8.8	−8.6	−8.7	−9.0	−11
UCKS	431	179	67.1	23.8	8.15	4.22	2.17	0.882	0.0886
	−31	−16	−8.1	−4.6	−3.6	−3.8	−4.3	−5.3	−8.8
'scaled UCKS'	415	174	65.8	23.4	8.01	4.14	2.13	0.863	0.0862
	−34	−18	−10	−6.2	−5.3	−5.5	−6.2	−7.3	−11
CKS	467	192	72.4	25.8	8.86	4.60	2.37	0.965	0.0974
	−25	−9.6	−0.9	3.3	4.8	4.8	4.5	3.7	0.2

[a] The second rows contain the percent errors with respect to the FCI exchange-dispersion energy (in the S^2 approximation). Energies are in μhartrees, distances are in bohr.

part of the exchange-dispersion energy. For larger distances, the differences are smaller than 2%, so they are of the same order of magnitude as some DF errors in Table 2, and it is difficult to attribute them unanimously to DF or cumulant sources of error. In order to clarify this point a non-DF calculation has been performed for a distance 5.6 bohr in a smaller DC77 basis, giving the value of the (noncumulant) CCSD exchange-dispersion energy (eq 35) equal to 2.2867 μhartree. For this distance, the $E_{\text{exch}-\text{disp}}^{(2)}$(FCI) energy in the S^2 approximation (2.2279 μhartree) is available from unpublished calculations, utilizing a SAPT program specific for two-electron monomers used also in refs 52, 63, and 71. A subtraction of these two numbers gives, therefore, a pure cumulant term for the helium dimer (eq 35). It turns out that for the van der Waals minimum region this contribution constitutes −2.6% of a total FCI exchange-dispersion energy, i.e., is small enough to be neglected (this difference is only 0.2% of the total interaction energy). Summarizing, the cumulant contribution for the helium dimer is significant only in highly repulsive regions of PES (where it contributes by ca. 25% to the total exchange-dispersion energy for He$_2$, see Table 3), while in all other regions this term is hidden within uncertainties caused by the DF approximation. Unfortunately, for many-electron monomers there is no reliable benchmark for the exchange-dispersion energy, so the issue of the importance of the cumulant contribution cannot be solved unambiguously at the moment. However, it is known for previous calculations that similar cumulant contributions for the first-order exchange[40] and second-order exchange-induction[42] energies contribute an unsubstantial fraction to the total energy correction. It can be, therefore, anticipated that the cumulant part of the exchange-dispersion energy should be small as well. It should be also noted that the calculation of the cumulant contribution for many-electron monomers would be very expensive, since six-index (five-index with DF) objects should have been stored for a number of frequencies on the one hand, and the numerical quadrature itself would scale with the eighth (seventh with DF) power of the molecular size on the other.

In Table 3 the UCHF, CHF and DFT exchange-dispersion energies for the helium dimer are also listed for a comparison. As it can be seen from this table, the errors of about 30% are common for the $E_{\text{exch}-\text{disp}}^{(2)}$(UCHF) energy. Errors become

about three times smaller if the CHF polarization propagators are used instead of the UCHF ones. On the other side, the exchange-dispersion energies obtained from the asymptotically corrected PBE0 functional and CKS propagators show similar errors relative to the $E_{\text{exch}-\text{disp}}^{(2)}$(FCI) exact values as the $E_{\text{exch}-\text{disp}}^{(2)}$(CCSD) results. Therefore, the conclusions can be made that the CKS propagators for helium reproduce rather well the exact propagators, if the asymptotically corrected PBE0 functional is used. On the other hand, the quality of the UCKS exchange-dispersion energies with the same functional is significantly worse. It should be stressed that the usage of the scaling formula (44) deteriorates the agreement with the FCI benchmark values in comparison to the unscaled UCKS energies for all distances considered.

In Table 4 the dispersion and exchange-dispersion energies for several van der Waals complexes are listed. These energies have been calculated with DF-CCSD, UCHF, CHF, and DFT approaches for monomers. A comparison of the results listed in Table 4 with the DF-CCSD values, serving as the benchmark, supports the conclusion reached for the helium dimer that the UCHF method underestimates substantially the exchange-dispersion energy (by even 38% for the He−HF complex). Unfortunately, the CHF approach does not help to the same extent, as in the case of He$_2$, although errors of the CHF corrections do become somewhat smaller.

Finally, let us examine the DFT exchange-dispersion energies. Again, three values of this correction listed in the table correspond to the UCKS, 'scaled UCKS' (see eq 44), and true CKS methods. It can be observed that the CKS results are remarkably close to the benchmark values in all cases, while errors in the UCKS and 'scaled UCKS' methods are much larger (with one exception of the Ne−Ar complex, for which the UCKS exchange-dispersion energy is close to the DF-CCSD benchmark). In particular, the 'scaled UCKS' approach does not represent any improvement over the UCKS approach, since usually the 'scaled UCKS' exchange-dispersion energy 'overshoots' the DF-CCSD value from the other side, although the errors are somewhat smaller in some cases. Summarizing, from three SAPT(DFT) approaches for a calculation of the exchange-dispersion only the method utilizing the CKS propagators can be recommended. If the CKS propagators are unavailable, one should rather use

Table 4. A Comparison of the Exchange-Dispersion Energies Computed with the DF-CCSD(3), UCHF, CHF, and UCKS, 'Scaled UCKS', and CKS Approaches for Monomers[a]

method	complex							
	Ne−Ar	Ar$_2$	(H$_2$O)$_2$	(HF)$_2$	(CO)$_2$	(N$_2$)$_2$	CO−H$_2$O	He−HF
				$E^{(2)}_{\text{exch−disp}}$				
DF-CCSD	0.0114	0.061	0.642	0.527	0.092	0.045	0.083	0.0053
UCHF	0.0085	0.055	0.495	0.381	0.064	0.029	0.059	0.0033
CHF	0.0093	0.059	0.527	0.400	0.072	0.035	0.063	0.0040
UCKS	0.0111	0.075	0.733	0.587	0.104	0.050	0.102	0.0051
'scaled UCKS'	0.0085	0.050	0.575	0.481	0.070	0.040	0.071	0.0045
CKS	0.0101	0.063	0.628	0.513	0.092	0.047	0.084	0.0054
				$E^{(2)}_{\text{disp}}$				
DF-CCSD	−0.267	−0.846	−3.651	−3.176	−1.048	−0.537	−1.013	−0.1663
UCHF	−0.244	−0.918	−3.116	−2.556	−0.967	−0.449	−0.910	−0.1261
CHF	−0.236	−0.815	−3.200	−2.673	−0.931	−0.477	−0.873	−0.1443
UCKS	−0.329	−1.270	−4.511	−3.758	−1.499	−0.662	−1.400	−0.1874
CKS	−0.253	−0.852	−3.543	−3.079	−1.016	−0.539	−0.976	−0.1650

[a] Dispersion energies are also listed for a comparison. Energies are in millihartrees.

UCKS propagators without trying to improve them by introducing some provisional scaling procedures.

4. Summary

A new formulation of the second-order exchange-dispersion energy in the single-exchange approximation has been presented. The exchange-dispersion energy is expressed through the formula involving frequency-dependent density-matrix susceptibilities and one-electron density matrices of monomers, in the spirit of the Longuet-Higgins formula for the second-order dispersion energy. An explicit cancellation of disconnected terms has been achieved through the introduction of the cumulant density susceptibilities. The new approach has been applied to the case of the CCSD polarization propagators for the main (noncumulant) part of the exchange-dispersion energy. In this way the exchange-dispersion energy with the inclusion of the intramonomer correlation effects has been obtained for the first time for many-electron systems. The density fitting technique has been applied to density-matrix susceptibilities in order to reduce the computational time. For the CCSD susceptibilities, the CPU time has been scaled down in this way from the eighth to seventh power in terms of the orbital basis size, thus allowing to perform benchmark calculations for a few-atom monomers. The relative error of density fitting for the case of the exchange-dispersion energy is somewhat larger than for the dispersion energy and sometimes constitutes a couple percent of the whole correction in the minimum and even about 10% or more for large distances. However, since the exchange-dispersion energy is usually much smaller than the dispersion energy, the error resulted from density fitting with respect to the sum of exchange-dispersion and dispersion energies is of order of a few tenths of percent. A comparison with benchmark DF-CCSD values has allowed to assess the quality of the three common SAPT(DFT) approaches for the calculation of the exchange-dispersion energy. It turns out that the CKS exchange-dispersion energy is usually in a very good agreement with DF-CCSD, while the quality of the UCKS energies is significantly worse in a majority of cases. Finally, the numerical investigation of the 'scaled UCKS' approach shows that this method should be abandoned in favor to the CKS exchange-dispersion energy.

5. Appendix

5.1. The Exchange-Induction Energy Expressed through Density-Matrix Susceptibilities. A computationally efficient way to express the second-order exchange-induction energy through monomer properties has been developed in ref 42 where the ground-state and first-order density matrices have been used for this purpose (with the effective electrostatic potential of the second monomer acting as a perturbing operator). From a theoretical point of view, it may be interesting to express this correction analogously to the induction energy, for which the following formula exists in the literature:[72,73]

$$E^{(2)}_{\text{ind}}(A \leftarrow B) = -\frac{1}{2} \int v_{\text{eff,B}}(1)\alpha_A(1, 1'|0)v_{\text{eff,B}}(1')d\tau_1 d\tau_1' \tag{47}$$

In eq 47, the effective electrostatic potential of the monomer B is used, which is defined as

$$v_{\text{eff,B}}(1) = v_B(1) + \int \frac{\rho_B(1'|1')}{r_{11'}}d\tau_1' \tag{48}$$

where in turn v_B is the one-electron potential of the monomer B. The formula for the $E^{(2)}_{\text{ind}}(B \leftarrow A)$ term is completely analogous. Repeating derivation steps shown in the first part of this paper, the exchange-induction energy in the S^2 approximation can be written in terms of the static density-matrix susceptibilities:

$$
\begin{aligned}
E_{\text{exch-ind}}^{(2)}(A \leftarrow B) \;=\; & \tfrac{1}{2}\int v(1,1')v_{\text{eff,B}}(3)\alpha_A(1|1';3|0)\rho_B(1'|1)d\tau_1 d\tau_1' d\tau_3 \\
& + \tfrac{1}{2}\int v(1,1')v_{\text{eff,B}}(3)\alpha_A(1|2';3|0)\mathscr{A}_{1'2'}^B \rho_B(1'|1)\rho_B(2'|1)d\tau_1 d\tau_1' d\tau_2' d\tau_3 \\
& + \tfrac{1}{2}\int v(1,1')v_{\text{eff,B}}(3)\mathscr{A}_{12}^{*A}\mathscr{A}_{12'}^A \alpha_A(1|1;3|0)\rho_A(2|1')\rho_B(1'|2)d\tau_1 d\tau_1' d\tau_2 d\tau_3 \\
& + \tfrac{1}{2}\int v(1,1')v_{\text{eff,B}}(3) \\
& \times \; [(\mathscr{A}_{12}^{*A}\mathscr{A}_{12'}^A\mathscr{A}_{1'2}^B - 1 - \pi_{12}^{*A}\pi_{1'2'}^A)\alpha_A(1|1;3|0)\rho_A(2|2')\rho_B(1'|1')\rho_B(2'|2) \\
& + \; \alpha_A(1|2';3|0)\Lambda_B(1'2'|1'2) + \lambda_A(12|12';3|0)\rho_B(1'|2) \\
& + \; \mathscr{A}_{12}^{*A}\mathscr{A}_{1'2'}^A \alpha_A(1|1;3|0)\rho_A(2|2')\Lambda_B(1'2'|1'2) \\
& + \; \lambda_A(12|12';3|0)\mathscr{A}_{1'2'}^B \rho_B(1'|1')\rho_B(2'|2) + \lambda_A(12|12';3|0)\Lambda_B(1'2'|1'2)]d\tau_1 d\tau_1' d\tau_2 d\tau_2' d\tau_3
\end{aligned}
\tag{49}
$$

where $\Lambda_B(12|1'2')$ denotes the two-electron density-matrix cumulant[74] of the monomer B, which for the CC case has been derived in ref 41. The first four terms inside the square brackets correspond to the noncumulant part of the exchange-induction energy. Equation 49 is interesting from a formal point of view, since it expresses the exchange-induction energy through quantities independent of another monomer. The calculation of this correction by using eq 49 would require a knowledge of the whole density-matrix susceptibility. Clearly, from the computational point of view, it is more convenient to calculate one response of the monomer with respect to the potential of another monomer and to use the resulting first-order density matrix for the calculation of the exchange-induction energy, than to calculate ca. N_{AO}^2 or N_{aux} responses to the $\hat{G}_q^p$ or $\hat{\chi}_K$ operators. However, eq 49 could be in principle an interesting alternative to the approach presented in ref 42 if it were sufficient to obtain monomer properties for SAPT in a purely monomer basis set.

5.2. Working Formula for the Exchange-Dispersion Energy. An implementation of eq 35 can be further simplified if one utilizes the generalized Coulomb and exchange matrices, introduced in ref 39. In the formula below the capital bold letters denote a matrix of $N_{\text{AO}} \times N_{\text{AO}}$ dimensions, the one-electron density matrices of monomers are denoted by the capital bold letters **A** and **B**, while **S** stands for the overlap matrix. The definitions of *generalized* right and left Coulomb matrices and *generalized* inner and outer exchange matrices for a general (not necessarily symmetric) matrix **Y** are listed below:

$$
\begin{aligned}
\overset{r}{\mathbf{J}}(\mathbf{Y})_r^s &= Y_{s'}^{r'}(rs|v|r's') \\
\overset{l}{\mathbf{J}}(\mathbf{Y})_{r'}^{s'} &= Y_s^r(rs|v|r's') \\
\overset{i}{\mathbf{K}}(\mathbf{Y})_r^{s'} &= Y_s^{r'}(rs|v|r's') \\
\overset{o}{\mathbf{K}}(\mathbf{Y})_{r'}^{s} &= Y_{s'}^{r}(rs|v|r's')
\end{aligned}
\tag{50}
$$

With these definitions the exchange-dispersion energy can be expressed as

$$
\begin{aligned}
E_{\text{exch-disp}}^{(2)}(n) \;=\; & 2X_{q_1 s_1}^{p_1 r_1}(p_1 s_1|v|r_1 q_1) \\
& + \; 2X_{q_1 s_2}^{p_1 r_2}\mathbf{S}_{r_2}^{q_1}\overset{r}{\mathbf{J}}(\mathbf{B})_{p_1}^{s_2} + 2X_{q_1 s_1}^{p_1 r_1}(\mathbf{SB})_{s_2}^{q_1}(p_1 s_2|v|r_1 s_1) - X_{q_1 s_1}^{p_1 r_2}\mathbf{S}_{r_2}^{q_1}\overset{i}{\mathbf{K}}(\mathbf{B})_{p_1}^{s_1} \\
& - \; X_{q_1 s_2}^{p_1 r_1}(\mathbf{SB})_{s_1}^{q_1}(p_1 s_2|v|r_1 s_1) \\
& + \; \left[2X_{q_2 s_1}^{p_2 r_1}\overset{l}{\mathbf{J}}(\mathbf{A})_{r_1}^{q_2}\mathbf{S}_{p_2}^{s_1} + 2X_{q_1 s_1}^{p_1 r_1}(\mathbf{SA})_{q_2}^{s_1}(p_1 q_1|v|r_1 q_2) - X_{q_1 s_1}^{p_2 r_1}\overset{o}{\mathbf{K}}(\mathbf{A})_{r_1}^{q_1}\mathbf{S}_{p_2}^{s_1}\right] \\
& - \; X_{q_2 s_1}^{p_1 r_1}(\mathbf{SA})_{q_1}^{s_1}(p_1 q_1|v|r_1 q_2) \\
& \quad + \; \tfrac{1}{2}\Big[4X_{q_2 s_1}^{p_2 r_1}(\mathbf{SBS})_{p_2}^{q_2}\overset{l}{\mathbf{J}}(\mathbf{A})_{r_1}^{s_1} + 4X_{q_1 s_2}^{p_1 r_2}(\mathbf{SAS})_{q_2}^{s_2}\overset{r}{\mathbf{J}}(\mathbf{B})_{p_1}^{q_1} \\
& \quad - \; 2X_{q_2 s_1}^{p_2 r_2}\mathbf{S}_{r_2}^{q_2}\overset{l}{\mathbf{J}}(\mathbf{A})_{r_1}^{r_1}(\mathbf{BS})_{p_2}^{r_1} - 2X_{q_2 s_2}^{p_2 r_1}\mathbf{S}_{p_2}^{s_2}(\mathbf{SB})_{s_1}^{q_2}\overset{l}{\mathbf{J}}(\mathbf{A})_{r_1}^{s_1} \\
& \quad - \; 2X_{q_1 s_1}^{p_1 r_2}(\mathbf{BSAS})_{r_2}^{r_1}(p_1 q_1|v|r_1 s_1) - 2X_{q_1 s_2}^{p_1 r_1}(\mathbf{SASB})_{s_1}^{s_2}(p_1 q_1|v|r_1 s_1) \\
& \quad - \; 2X_{q_1 s_2}^{p_2 r_2}\mathbf{S}_{p_2}^{s_2}\overset{r}{\mathbf{J}}(\mathbf{B})_{p_1}^{q_1}(\mathbf{AS})_{r_2}^{p_1} - 2X_{q_1 s_1}^{p_2 r_1}(\mathbf{ASBS})_{p_2}^{p_1}(p_1 q_1|v|r_1 s_1) \\
& \quad - \; 2X_{q_2 s_2}^{p_1 r_2}(\mathbf{SA})_{q_1}^{s_2}\overset{r}{\mathbf{J}}(\mathbf{B})_{p_1}^{q_1}\mathbf{S}_{r_2}^{q_2} - 2X_{q_2 s_1}^{p_1 r_1}(\mathbf{SBSA})_{q_1}^{q_2}(p_1 q_1|v|r_1 s_1) \\
& \quad + \; X_{q_1 s_1}^{p_2 r_2}(\mathbf{AS})_{r_2}^{p_1}(\mathbf{BS})_{p_2}^{r_1}(p_1 q_1|v|r_1 s_1) + X_{q_1 s_2}^{p_2 r_1}\overset{o}{\mathbf{K}}(\mathbf{ASB})_{r_1}^{q_1}\mathbf{S}_{p_2}^{s_2} \\
& \quad + \; X_{q_2 s_1}^{p_1 r_2}\mathbf{S}_{r_2}^{q_2}\overset{i}{\mathbf{K}}(\mathbf{BSA})_{p_1}^{s_1} + X_{q_2 s_2}^{p_1 r_1}(\mathbf{SB})_{s_1}^{q_2}(\mathbf{SA})_{q_1}^{s_2}(p_1 q_1|v|r_1 s_1)\Big]
\end{aligned}
\tag{51}
$$

Exchange-Dispersion Energy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2677**

The quantities like products of the overlap and density matrices or the generalized Coulomb and exchange matrices can be calculated in advance and stored on disk. Unfortunately, not all terms can be factorized in this way, so still the overall scaling of eq 51 remains (N_{AO}^5), although with a lower prefactor. As already noted, the scaling of eq 51 is of a limited importance in view of high costs of the DF-CCSD propagators.

5.3. A Comparison with the Formula for the UCHF Exchange-Dispersion Energy. In order to find a correspondence between the new formula (eq 35) and the UCHF exchange-dispersion energy (eq 108 of ref 7) let us consider the first term of eq 35, in which the UCHF propagators have been used. If Ψ^0 is set to Φ (Hartree−Fock determinant) and Ψ^K are approximated by singly-excited determinants $\Psi^K = \Phi(i \rightarrow a)$, then the propagator formula (see eq 28) takes the form:

$$
\begin{aligned}
\Pi_{q_1 q_2}^{p_1 p_2}(i\omega) &= 2\sum_{ia}\langle\Phi|\hat{e}_{q_1}^{p_1}\Phi(i\rightarrow a)\rangle\langle\Phi(i\rightarrow a)|\hat{e}_{q_2}^{p_2}\Phi\rangle\frac{\Delta_{ai}}{(\Delta_{ai})^2+\omega^2} \\
&= 2\sum_{ia}\delta_i^{p_1}\delta_{q_1}^a\delta_a^{p_2}\delta_{q_2}^i\frac{\Delta_{ai}}{(\Delta_{ai})^2+\omega^2}
\end{aligned}
\tag{52}
$$

where $\Delta_{ai} = \varepsilon_a - \varepsilon_i$ is a difference of orbital energies for spinorbitals ϕ_a and ϕ_i (a,b and i,j denote virtual and occupied indices, respectively, relative to Φ). If eq 52 is inserted into the first term of eq 35, then the following expression is obtained

$$
\begin{aligned}
&\frac{1}{2\pi}\int_0^\infty {}^A\Pi_{q_1 q_2}^{p_1 p_2}(i\omega){}^B\Pi_{s_1 s_2}^{r_1 r_2}(i\omega)d\omega(p_1 s_1|v|r_1 q_1)(p_2 q_2|r_2 s_2) \\
&= \sum_{ia\in A}\sum_{jb\in B}\frac{2}{\pi}\int_0^\infty\frac{\Delta_{ai}}{(\Delta_{ai})^2+\omega^2}\frac{\Delta_{bj}}{(\Delta_{bj})^2+\omega^2}d\omega(ib|v|ja)(ai|bj) \\
&= \sum_{ia\in A}\sum_{jb\in B}\frac{1}{\Delta_{ai}+\Delta_{bj}}(ai|bj)(ib|v|ja) = -t_{ab}^{ij}(ib|v|ja)
\end{aligned}
\tag{53}
$$

which, after the summation over spins, gives the first term of eq 108 in ref 7. Note that this paper contains several misprints corrected recently in ref 75.

References

(1) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.

(2) Szalewicz, K.; Jeziorski, B. Symmetry-Adapted Perturbation Theory of Intermolecular Interactions. In *Molecular Interactions−from van der Waals to Strongly Bound Complexes*; Schreiner, S., Ed.; Wiley: New York, 1997; pp 3−43.

(3) Szalewicz, K.; Patkowski, K.; Jeziorski, B. *Struct. Bonding (Berlin)* **2005**, *116*, 43.

(4) Moszynski, R.; Jeziorski, B.; Ratkiewicz, A.; Rybak, S. *J. Chem. Phys.* **1993**, *99*, 8856.

(5) Szalewicz, K.; Jeziorski, B. *Mol. Phys.* **1979**, *38*, 191.

(6) Jeziorski, B.; Moszynski, R.; Rybak, S.; Szalewicz, K. Many-Body Theory of van der Waals Interactions. In *Many-Body Methods in Quantum Chemistry*; Kaldor, U., Ed.; Springer: New York, 1989; Vol. *52 of Lecture Notes in Chemistry*; pp 65−94.

(7) Rybak, S.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1991**, *95*, 6576.

(8) Moszynski, R.; Jeziorski, B.; Szalewicz, K. *Int. J. Quantum Chem.* **1993**, *45*, 409.

(9) Moszynski, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1994**, *100*, 1312.

(10) Moszynski, R.; Cybulski, S. M.; Chalasinski, G. *J. Chem. Phys.* **1994**, *100*, 4998.

(11) Moszynski, R.; Jeziorski, B.; Rybak, S.; Szalewicz, K.; Williams, H. L. *J. Chem. Phys.* **1994**, *100*, 5080.

(12) Williams, H. L.; Szalewicz, K.; Moszynski, R.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 4586.

(13) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorska, M.; Jeziorski, B.; Kucharski, S. A.; Lotrich, V. F.; Misquitta, A. J.; Moszyński, R.; Patkowski, K.; Podeszwa, R.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wheatley, R. J.; Wormer, P. E. S.; Żuchowski, P. S. *SAPT2008: An Ab Initio Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies*, University of Delaware and University of Warsaw, 2008; http://www.physics.udel.edu/~szalewic/SAPT/SAPT.html.

(14) Heijmen, T. G. A.; Moszynski, R.; Wormer, P. E. S.; van der Avoird, A. *Mol. Phys.* **1996**, *89*, 81.

(15) Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **2000**, *113*, 6687.

(16) Korona, T.; Moszynski, R.; Thibault, F.; Launay, J.-M.; Bussery-Honvault, B.; Boissoles, J.; Wormer, P. E. S. *J. Chem. Phys.* **2001**, *115*, 3074.

(17) Bussery-Honvault, B.; Launay, J.-M.; Moszynski, R. *Phys. Rev. A: At. Mol. Opt. Phys.* **2003**, *68*, 032718.

(18) Jaszunski, M.; McWeeny, R. *Mol. Phys.* **1985**, *55*, 1275.

(19) Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 646.

(20) Jansen, G.; Hesselmann, A. *J. Phys. Chem. A* **2001**, *105*, 11156.

(21) Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301.

(22) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(23) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

(24) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464.

(25) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319.

(26) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778.

(27) Hesselmann, A.; Jansen, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5010.

(28) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 4993.

(29) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

(30) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Am. Chem. Soc.* **2006**, *128*, 11730.

(31) Podeszwa, R.; Szalewicz, K. *Chem. Phys. Lett.* **2005**, *412*, 488.

(32) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Chem. Theory Comput.* **2006**, *2*, 400.

(33) Podeszwa, R.; Bukowski, R.; Rice, B. M.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5561.

(34) Fiethen, A.; Jansen, G.; Hesselmann, A.; Schütz, M. *J. Am. Chem. Soc.* **2008**, *130*, 1802.

(35) Korona, T.; Hesselmann, A.; Dodziuk, H. *J. Chem. Theor. Comp.* **2009**, *5*, 1585.

(36) Korona, T.; Jeziorski, B.; Moszynski, R. *Mol. Phys.* **2002**, *100*, 1723.

(37) Korona, T.; Jeziorski, B. *J. Chem. Phys.* **2006**, *125*, 184109.

(38) Korona, T. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6004.

(39) Korona, T.; Jeziorski, B. *J. Chem. Phys.* **2008**, *128*, 144107.

(40) Korona, T. *J. Chem. Phys.* **2008**, *122*, 224104.

(41) Korona, T. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5698.

(42) Korona, T. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6509.

(43) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 36.

(44) Bulski, M.; Chalasinski, G.; Jeziorski, B. *Theor. Chim. Acta* **1979**, *52*, 93.

(45) Chalasinski, G.; Jeziorski, B. *Mol. Phys.* **1976**, *32*, 81.

(46) Chalasinski, G.; Jeziorski, B. *Theor. Chim. Acta* **1977**, *46*, 477.

(47) Longuet-Higgins, H. C. *Faraday Discuss. Chem. Soc.* **1965**, *40*, 7.

(48) Paldus, J.; Jeziorski, B. *Theor. Chim. Acta* **1988**, *73*, 81.

(49) Moszynski, R.; Żuchowski, P. S.; Jeziorski, B. *Collect. Czech. Chem. Commun.* **2005**, *70*, 1109.

(50) Jeziorski, B.; Moszynski, R. *Int. J. Quantum Chem.* **1993**, *48*, 161.

(51) Korona, T.; Przybytek, M.; Jeziorski, B. *Mol. Phys.* **2006**, *104*, 2303.

(52) Korona, T.; Williams, H. L.; Bukowski, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 5109.

(53) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 7374.

(54) Paldus, J. In *Theoretical Chemistry Advances and Perspectives*; Eyring, H., Henderson, D., Eds.; Academic Press: New York, 1976; p 177.

(55) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO, version 2008.2, a package of ab initio programs*, 2008, see http://www.molpro.net.

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

(57) Schütz, M.; Manby, F. R. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3349.

(58) Dunning, Jr., T. H.; Hay, P. J. In *Methods of Electronic Structure Theory*; Schaefer H. F., III, Ed.; Plenum Press: New York, 1977; Vol 2.

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

(60) Kendall, R.; Dunning, T., Jr.; Harrison, R. *J. Chem. Phys.* **1992**, *96*, 6769.

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

(62) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59.

(63) Williams, H. L.; Korona, T.; Bukowski, R.; Szalewicz, K.; Jeziorski, B. *Chem. Phys. Lett.* **1996**, *262*, 431.

(64) Wada, A.; Kanamori, H.; Iwata, S. *J. Chem. Phys.* **1998**, *109*, 9434.

(65) Rode, M.; Sadlej, J.; Moszynski, R.; Wormer, P. E. S.; van der Avoird, A. *Chem. Phys. Lett.* **1999**, *314*, 326.

(66) Milet, A.; Korona, T.; Moszynski, R.; Kochanski, E. *J. Chem. Phys.* **1999**, *111*, 7727.

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

(68) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(69) Grüning, M.; Gritsenko, O. V.; van Gisergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2001**, *114*, 652.

(70) Computational Chemistry Comparison and Benchmark DataBase, http://cccbdb.nist.gov; (accessed July 1, 2008).

(71) Korona, T.; Moszynski, R.; Jeziorski, B. *J. Chem. Phys.* **1996**, *105*, 8178.

(72) Ángýan, J. G.; Jansen, G.; Loos, M.; Hättig, C.; Hess, B. A. *Chem. Phys. Lett.* **1994**, *219*, 267.

(73) Jeziorski, B.; Szalewicz, K. Symmetry-Adapted Perturbation Theory. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; Wiley: New York, 2003; Vol. 3, p 232.

(74) Kutzelnigg, W.; Mukherjee, D. *J. Chem. Phys.* **1997**, *107*, 432.

(75) Bukowski, R.; Cencek, W.; Patkowski, K.; Jankowski, P.; Jeziorska, M.; Kolaski, M.; Szalewicz, K. *Mol. Phys.* **2006**, *104*, 2241.

JCTC Journal of Chemical Theory and Computation

Combined Jahn−Teller and Pseudo-Jahn−Teller Effect in the CO₃ Molecule: A Seven-State Six-Mode Problem

Yang Liu, Isaac B. Bersuker, Wenli Zou, and James E. Boggs*

Institute for Theoretical Chemistry, Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, Texas 78712-0165

Abstract: A complicated problem of seven electronic states in four terms, $^1A_1'$, $^1E''$, $1^1E'$, and $2^1E'$, interacting with six vibrational modes, a_1', a_2'', e', and e', was solved to take into account the combined two-mode Jahn−Teller (JT) plus two-mode pseudo JT effects and rationalize the electronic structure of the CO₃ molecule. The JT and first-order pseudo JT effects in the E″ state are separated from the rest of the problem by symmetry; they do not influence the ground state properties. In the remaining five-state five-mode problem including the ground state, $(A_1' + 1E' + 2E') \otimes (a_1' + e' + e')$, the JT two-mode problem is reduced to the one-mode one by means of coordinate transformations. Several high-level ab initio calculations including all of the five states confirm the previously found coexistence of a central minimum of D_{3h} symmetry and three equivalent minima with a distorted geometry of C_{2v} symmetry in the ground state; the barrier between them is rather small, 0.2−0.3 eV, but large enough to distinguish them spectroscopically. Harmonic vibrational frequencies of the two configurations near the minima of the adiabatic potential energy surface are also evaluated. The calculations show how the distorted configurations are produced by the JT effect in one of the excited E states, similar to a previous finding in O₃. Numerical data of ab initio calculations yield also the effective vibronic and primary force constants for all of the terms. An electronic structure problem of this complexity including a reduction of the two-mode problem to one mode with full interpretation of the origin of coexisting different geometries as due to the JTE in the excited state is presented here for the first time.

Introduction

The carbon trioxide CO₃ molecule is of significant interest due to its potential role in the atmospheric chemistry of Earth and Mars.[1−9] The necessity to characterize the CO₃ molecule emerged first from the data on the photolysis of ozone−carbon dioxide ices at 77 K.[2] Attention to this system was renewed recently by the studies of kinetics of the reaction of carbon dioxide CO₂ with oxygen.[1−8] It was proved that CO₃ is an important intermediate in the quenching of the electronically excited oxygen atoms O(^{1}D) by CO₂ and the ^{18}O isotope enrichment of stratospheric carbon dioxide. The O(^{1}D) atom is initially generated by the photolysis of stratospheric ozone; it then collides with CO₂ to form CO₃, followed by dissociation to CO₂ and the O atom in the ground ^{3}P or excited ^{1}D electronic state via oxygen quenching or an isotope exchange mechanism.

Quite a number of experimental and theoretical investigations have been devoted to the structure of the CO₃ molecule. Ung and Schiff[10] in 1966 postulated the existence of the CO₃ species in the gas phase without its explicit detection. Later, Moll et al.[2] and Jacox and Milligan[4] investigated the infrared spectrum and derived a planar C_{2v} structure at low temperatures. Earlier experiments predicted that CO₃ in the C_{2v} configuration is produced in the reaction of CO₂ + O(^{1}D)[11]. The latest experimental exploration of the structure of CO₃ was carried out by Jamieson and co-workers in 2006.[12] It was the first spectroscopic detection and identification of the D_{3h} configuration of CO₃.

* To whom correspondence should be addressed. E-mail: james.boggs@mail.utexas.edu.

In 1987, Van de Guchte et al.[13] performed self-consistent field (SCF) calculations on CO_3 and predicted a stable D_{3h} structure with lower energy than in the C_{2v} geometry. This calculation is likely to be too simple to give a reliable conclusion, but it inspired further interest in elucidating the geometric and electronic structure of CO_3. Castro and Sylvio[14] used many-body perturbation theory and couple-cluster calculations to confirm that the C_{2v} and D_{3h} isomers coexist, but the C_{2v} structure lies about 18 kJ/mol lower than the D_{3h} isomer.

In recent years, higher-level computational methods have been used extensively to explore the structure of the CO_3 molecule. In 2004, on the basis of the complete active space self-consistent field (CASSCF) optimized geometries and internal contracted multireference configuration interaction with Davidson's cluster correction (MRCI + Q) single-point energies, Mebel et al.[15] and Bennett et al.[8] predicted that the C_{2v} and D_{3h} configurations have similar energies (the energy difference is about 0.1 kcal/mol) and are separated by a transition state with a barrier of about 4.4 kcal/mol. In 2007, Kowalczyk and Krylov[16] performed a series of coupled cluster calculations (CCSD, CCSD(T), CCSDT, EOM-CCSD, and EOM-EE(2,3)) to study the electronic structure of CO_3. They were the first to involve the influence of the excited state via the pseudo Jahn−Teller effect (JTE) in the formation of the ground state geometry of this molecule (for other systems, this procedure has been previously involved many times[17,18]), but they took into account only one excited state in a one-mode treatment. In the calculations of this paper,[16] the D_{3h} minimum disappears at higher levels of correlation treatment and basis set, and only the C_{2v} minimum remains. The authors explained the disappearance of the high-symmetry equilibrium configuration by pointing out that in previous multireference studies dynamical correlation was not included in the geometry optimization. More recently, Qin and Soo[19] used the DFT method to support the conclusion of coexistence of the D_{3h} and C_{2v} geometries.

Despite these theoretical studies performed on CO_3, the main question of the ground state configuration and geometry, D_{3h} versus C_{2v}, as well as expected vibrational frequencies at the minima, remains controversial, and no full discussion of the origin of this complicated adiabatic potential energy surface (APES) is given. The presence of a C_{2v} minimum with a symmetry lower than the highest one, D_{3h}, in the nondegenerate ground state indicates that there is an essential implication of the JTE involving excited electronic states.[17,18] This means that to rationalize the results of ab initio calculations one has to take into account the mixing of the ground and all relevant excited states via the JTE and pseudo JTE (PJTE), which seems to be crucial in this problem. In any full consideration of this system, the three close-in-energy excited E states, each of which is potentially subject to JTE and PJT mixing with the others and with the ground state via the six active vibrations, must be taken into account. This amounts to a combined JTE plus PJTE seven-state six-mode problem, which can be reduced by symmetry to three simpler problems, the main of which is the five-state five-mode one that includes the ground state. The present paper reports a solution of these problems which rationalizes the electronic structure of CO_3 and gives a clear physical picture of the origin of its two coexisting configurations in the ground state.

Formally, the APES obtained in this paper may serve as a basis for evaluation of the full rovibronic spectrum of CO_3. This is a separate and very difficult problem; we do not know of any precedent rovibronic solutions of this complexity. Partial solutions were discussed by Stanton[20] for the radical NO_3, which looks similar to CO_3 but is essentially different.

The JTE and PJTE in CO_3

According to the one-point calculations of the CO_3 molecule in the high-symmetry configuration D_{3h} in this and previous works, its ground state is $^1A_1'$, and there are three spin-singlet excited electronic terms, E″, 1E′, and 2E′, that we found located at about, respectively, $\Delta = 0.97$ eV, $\Delta_1 = 2.65$ eV, and $\Delta_2 = 2.81$ eV above the ground state, and six vibrational modes: totally symmetric a_1', out-of-plane a_2'', bend e' (denoted below as $1e'$), and CO stretch e' (denoted below as $2e'$). In principle, all of these electronic terms and vibrations are involved in the formation of the APES of this molecule and its properties.

The first simplification of the problem can be obtained based on the fact that the E″ term does not mix with the ground state A_1' (in first order of the displacements) as there are no e'' vibrations in CO_3, and it mixes with the higher excited 1E′ and 2E′ terms via the out-of-plane displacements a_2'', which are not active in the remaining five-state five-mode problem. This allows us to separate the main combined $(A_1' + 1E' + 2E') \otimes (a_1' + 1e' + 2e')$ two-mode JT plus two-mode PJT problem[17] from the pure excited state JT E″ $\otimes (a_1' + 1e' + 2e')$ and PJT (E″ + 1E′ + 2E′) $\otimes a_2''$ problems. The vibronic coupling effects in the latter two problems are rather weak and not significant (see below).

The 5×5 secular matrix of the main problem contains a large number of constants including the JT vibronic coupling constants of each of the electronic terms to each of the vibrations and the PJT coupling constants between them, as well as the primary force constants and quadratic coupling constants. Fortunately, the problem can be simplified employing the results of the general theory of the multimode JTE (see in ref 17, sections 3.5 and 5.5; ref 21, and ref 22, as well as ref 23). According to the theory, the JT two-mode problem can be reduced to the ideal (one-mode) one by means of transformations in the coordinate system. In the transformed coordinates, only one e mode is JT-active (the *interaction mode*), the other one remaining non-JT-active (meaning harmonic and nonshifted), so the problem becomes $(A_1' + 1E' + 2E') \otimes (a_1' + e')$, and the JT plus PJT secular matrix is considerably simplified. The interaction mode stands for the total distortion produced by the nuclear displacements in the two modes, which in an appropriately chosen coordinate system is described by one minimum of the APES. Although the interaction mode is a function of the initial modes, and the reduced number of vibronic coupling constants is a combination of the larger number of initial constants, the problem is still essentially simplified because, as shown below, the interaction mode can be obtained from the ab initio calculations.
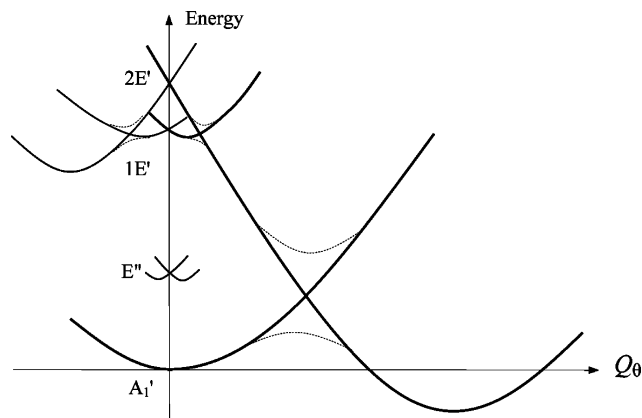
Figure 1. Schematic diagram of energy level of the ground state A_1' and four excited state ($1E'$ and $2E'$ in D_{3h} symmetry) dependence on the Q_θ coordinate of the interaction mode when their PJT mixing is ignored ($F_{ij} = 0$). The order of magnitude of the JT vibronic coupling constants is taken approximately in accordance with the ab initio calculations.

To get preliminary qualitative hints on where to find the interaction mode and what should be expected from the numerical calculations, some initial estimates based on the general theory[17] and previous experience[18] may be useful. In particular, if we ignore the PJT interaction between the three terms and take the JT coupling constants approximately as they appear in the ab initio calculations, we obtain for the cross section $Q_\varepsilon = 0$ the energy level dependence on Q_θ shown in Figure 1. In fact, to get this picture qualitatively, we should assume that the JTE in at least one of the exited E states is sufficiently strong so that the JT stabilization energy is larger than the energy gap to the ground state. This is just the picture we got in solving the problem of O_3.[18]

In Figure 1, the JTE splitting in the two excited E states and the harmonic curve of the ground state are shown. We know qualitatively how these curves change when the PJTE interaction is included; it is indicated by the dashed lines. An important feature in this energy level scheme is that the JTE in the excited state strongly influences the ground state APES, resulting, as shown in the ab initio calculations below, in the formation of an additional minimum with lower (C_{2v}) symmetry. There are three equivalent directions of the JTE distortions in the (Q_θ, Q_ε) space, and hence, if an additional C_{2v} minimum is formed under the influence of the excited state, there will be two more of them, forming three additional equivalent minima in the space of the interaction mode.

Following the qualitative picture of Figure 1 (which is confirmed below by the ab initio calculations), we see that the direction from the central minimum (if any) to the distorted configuration C_{2v} is determined by the direction of the JT distortion in the excited E states, more precisely, in the direction of the combined two-mode (interaction mode) distortion plus a possible shift of the a_1' displacements. To have an idea how the interaction mode depends on the initial modes, consider first the $2E'$ term with the strongest JTE, the $2E' \otimes (1e' + 2e' + a_1')$ problem, which generally produces two kind of distortions, one of which is given in symmetrized displacements ($q_{1\theta}$, $q_{1\varepsilon}$, q_a) of the $1e'$ mode, the other in the coordinates ($q_{2\theta}$, $q_{2\varepsilon}$, q_a) of the second mode $2e'$. Obviously, the real displacements of individual atoms

are some vector sums of their displacements in each of these modes. The three modes have different frequencies.

The process of reducing the two-mode problem to a one-mode one is to perform first a scale transformation, $(q_i')^2 = K_i q_i^2$, that makes the elastic term isotropic (with the same force constant $K = 1$ for all q_i' coordinates)[22] and then an orthogonal transformation (rotation) for $q_{i\theta}'$ and $q_{i\varepsilon}'$ and a shift for q_a' of the coordinate systems ($q_{i\theta}'$, $q_{i\varepsilon}'$, q_a') to new ones ($Q_{i\theta}$, $Q_{i\varepsilon}$, Q_a) in such a way that one of the modes becomes JT-inactive (has no linear terms in the Hamiltonian), the other one being accordingly modified. The transformation of the two-mode problem to a one-mode one with quadratic terms included was carried out recently by Polinger[23] (in previous works, only linear displacements were taken into account). In this transformation, the coupling to the totally symmetric coordinates can be separated by means of a shift along the Q_a axis. Performing the 2×2 transformation in the remaining (Q_θ, Q_ε) space, we get

$$Q_\theta = c_1 q_{1\theta} + c_2 q_{2\theta}$$
$$Q_\varepsilon = d_1 q_{1\varepsilon} + d_2 q_{2\varepsilon} \qquad (1)$$

where the coefficients of this transformation are combinations of linear and quadratic vibronic coupling constants and force constants of all three modes.[23] Their exact numerical evaluation is not the goal of our calculation, but in further exploration of the excited states, this may be necessary. The main role of the interaction mode is to allow us to reformulate the problem in terms of $(A_1' + 1E' + 2E') \otimes (a_1' + e')$ with one e mode instead of two and to determine the effective JT parameters of this interaction mode (meaning the parameters of the APES of the excited state) by comparison with the numerical data. Numerically, the constants in eq 1 will also be affected by the influence of the JTE on the $1E'$ term, but this influence is insignificant. The JTE on this term, as shown in the calculations below, is very small, and the contribution of its quadratic terms (which only contribute to the warping of the Mexican hat) is negligible (anyhow, it just changes the expressions for the coefficients in eq 1).

For the direction from the D_{3h} minimum to the expected C_{2v} minimum which coincides with the similar direction to the corresponding minimum of the JTE $2E' \otimes (a_1' + e')$ in the excited $2E'$ term (Figure 1), we should take into account that the minimum on the Q_θ axis (where $Q_\varepsilon = 0$) has also a shift component along the Q_a axis. This follows directly from the general solution of the quadratic $E \otimes (a_1' + e')$ problem[17] and is confirmed below in the ab initio calculations of the APES along the Q_a coordinate (see Figure 7). Therefore, in the space of the interaction mode, the direction from D_{3h} to C_{2v} to be matched with the numerical calculations below is

$$Q_\theta = c_1 q_{1\theta} + c_2 q_{2\theta} + c q_a \qquad (2)$$

With only one active e' mode taken into account and in the space of its coordinates, we can write the effective matrix equation for the potential energy of the system (the APES) of the $(A_1' + 1E' + 2E') \otimes (a_1' + e')$ problem as follows (**I** is a unit matrix 5×5):

$$\mathbf{W} - E\mathbf{I} = 0 \qquad (3)$$

where

$$W =$$

$$\begin{bmatrix}
W_1 & F_{12}Q_\theta & F_{12}Q_\varepsilon & F_{13}Q_\theta & F_{13}Q_\varepsilon \\
F_{12}Q_\theta & \begin{array}{c} W_2 - F_{22}Q_\theta + N_{22} + F_2^a Q_a + \\ F_2^{ae}Q_aQ_\theta + \Delta_1 \end{array} & F_{22}Q_\varepsilon + 2G_{22}Q_\varepsilon Q_\theta + F_2^{ae}Q_aQ_\varepsilon & F_{23}Q_\theta + F_{23}^a Q_a & F_{23}Q_\varepsilon \\
F_{12}Q_\varepsilon & F_{22}Q_\varepsilon + 2G_{22}Q_\varepsilon Q_\theta + F_2^{ae}Q_aQ_\varepsilon & \begin{array}{c} W_2 + F_{22}Q_\theta - N_{22} + F_2^a Q_a - \\ F_2^{ae}Q_aQ_\theta + \Delta_1 \end{array} & F_{23}Q_\varepsilon & F_{23}Q_\theta + F_{23}^a Q_a \\
F_{13}Q_\theta & F_{23}Q_\theta + F_{23}^a Q_a & F_{23}Q_\varepsilon & \begin{array}{c} W_3 - F_{33}Q_\theta + N_{33} + F_3^a Q_a + \\ F_3^{ae}Q_aQ_\theta + \Delta_2 \end{array} & F_{33}Q_\varepsilon + 2G_{33}Q_\varepsilon Q_\theta + F_3^{ae}Q_aQ_\varepsilon \\
F_{13}Q_\varepsilon & F_{23}Q_\varepsilon & F_{23}Q_\theta + F_{23}^a Q_a & F_{33}Q_\varepsilon + 2G_{33}Q_\varepsilon Q_\theta + F_3^{ae}Q_aQ_\varepsilon & \begin{array}{c} W_3 + F_{33}Q_\theta - N_{33} + F_3^a Q_a - \\ F_3^{ae}Q_aQ_\theta + \Delta_2 \end{array}
\end{bmatrix}$$

In the formation of this equation, the wave functions of the two E_i terms are denoted by $|i\theta\rangle$ and $|i\varepsilon\rangle$, $i = 2$ and 3, and that of the ground state is $|a\rangle$. The 5×5 secular matrix (eq 3) contains the effective constants of vibronic coupling to e' displacements, with linear F_{ii} ($i = 2, 3$) and quadratic G_{ii} ($i = 2, 3$), and three constants F_{ij} ($ij = 12, 13$, and 23) of pseudo JT coupling between the terms by e' displacements, as well as the diagonal quadratic coupling terms $N_{ii} = G_{ii}(Q_\theta^2 - Q_\varepsilon^2)$ and elastic energies $W_i = (1/2)K_i(Q_\theta^2 + Q_\varepsilon^2) + (1/2)K_i^a Q_a^2$, $i = 1–3$, and the constants of the vibronic coupling of the three electronic terms to Q_a displacements F_i^a ($i = 2, 3$) and one constant F_{23}^a of the coupling between the two E terms. The F_i^{ae} ($i = 2, 3$) are the quadratic constants for interactions between the a_1' and e' modes.

We emphasize that because of the transformations above the relation between these constants and the initial JT constants of coupling to the two e' modes are complicated and in fact unnecessary, as the latter are unobservable. The APES matrix (eq 3) should be regarded as the *effective potential energy matrix*, which describes the real surface to be obtained in the ab initio calculations below. The parameters in this matrix are thus *effective* coupling constants that describe the APES. Some of these constants are not directly related to observable properties (e.g., the primary force constants K_i for JT modes[17]). The derivations and discussion above show how the real APES is formed by complicated JT interactions between the electronic states and vibrational modes in the D_{3h} configuration, especially, why we can consider only one JT vibrational mode instead of two. After accepting this conclusion, we can ignore eqs 1 and 2 and start further consideration from eq 3.

As mentioned above, the electronic structure of CO_3 is somewhat similar to that in the ozone molecule O_3 where the three minima of the obtuse triangular (C_{2v}) configuration are formed under the strong influence of the JTE in the excited E state,[18] and there is a shallower minimum with the D_{3h} configuration. But compared with O_3, the CO_3 molecule is much more complicated because of the two-mode problem and the presence of three excited close-in-energy E terms in the high-symmetry configuration, with two of them strongly influencing the ground state. Among four-atomic molecules, CO_3 is the most complicated system studied in this respect.

Ab Initio Calculations

In all of the ab initio calculations for this problem, Dunning's correlation consistent basis set cc-pVTZ[24] is employed. The active space is composed of the extensively used 16 electrons and 13 active orbitals (i.e., all valence electrons and orbitals with the exception of 2s lone pairs of oxygen atoms) denoted as (16/13). Another smaller active space is (16/10), which excludes the three empty orbitals corresponding to the three antibonding σ_{CO}*'s based on (16/13). The inactive orbitals (1s of C and 1s and 2s of O) were kept frozen temporarily only when the smaller active space (16/10) was used.

The geometry optimizations and frequency calculations were performed by single-state CASSCF (SS-CASSCF)[25] and MRCI+Q methods[26] (Tables 1 and 2). In addition, state-average CASSCF (SA-CASSCF)[27] calculations were carried out to optimize the stationary points and scan the potential energy curves. On the basis of the SS-CASSCF or SA-CASSCF geometry, the MRCI+Q single-point energies are also obtained to account for the dynamical correlation. Considering the computational efficiency, only reference configurations with weights larger than 1×10^{-4} were selected to be correlated in the MRCI+Q calculations with 16/13 active space.

Besides the multireference methods above, two high-level single-reference methods, that is, Brueckner coupled-cluster doubles with perturbative triples (BCCD(T))[28] and coupled-cluster involving singles and doubles with perturbative triples and quadruples (CCSD(TQ)),[29] were used to optimize the structures and calculate the frequencies at the minima.

Table 1. Optimized Geometries (Bond Lengths in Ångstroms and Angles in Degrees), Energy Differences ΔE (in eV), and Total Energies (in au) at the D_{3h} Minimum of the CO_3 Molecule Obtained by Different Computational Methods

		C_{2v}				
methods	D_{3h}, $R(CO)$	$R(CO^1)$	$R(CO^2)$	$\angle O^1CO^2$	ΔE ($E_{C2v} - E_{D3h}$)	total energy (D_{3h})
MRCI+Q (16/10)	1.2550	1.1718	1.3252	142.2	−0.11	−263.27351667
MRCI+Q//SS-CASSCF(16/13)	1.2637	1.1739	1.3380	142.1	−0.10	−263.28928534
MRCI+Q//SA-CASSCF(16/13)	1.2635	1.1727	1.3428	142.2	−0.08	−263.28907391
BCCD(T)	1.2597	1.1758	1.3333	142.6	−0.23	−263.29442881
CCSD(TQ)	1.2570	1.1745	1.3324	142.4	−0.28	−263.29093183

A Seven-State Six-Mode Problem

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2683**

Table 2. Frequencies of CO_3 in cm^{-1} at the Minima Obtained with Different Methods

mode (symmetry)\methods	MRCI+Q (16/10)	SS-CASSCF(16/13)	BCCD(T)	CCSD(TQ)	CCSD(T)[a]
		D_{3h}			
bend (e')	460	440	455	466	−402
OPLA (a_2'')	769	761	762	772	776
CO stretch (a_1')	1124	1084	1094	1113	1105
CO stretch (e')	1647	1182	1316	1274	1093
		C_{2v}			
bend (b_2)	591	580	567	575	575
bend (a_1)	604	578	612	597	606
OPLA (b_1)	694	682	672	675	683
CO stretch (b_2)	1049	1017	1006	1017	1007
OCO stretch (a_1)	1122	1075	1093	1096	1099
CO stretch (a_1)	2100	2071	2066	2073	2078

[a] Ref 16.
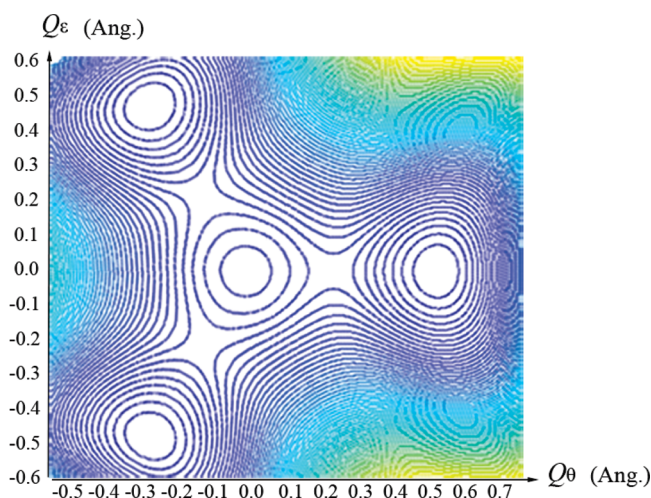
Figure 2. Equipotential curves on the APES of the ground state of CO_3 in the cross section of the (Q_θ, Q_ε) plane of the interaction mode obtained by the BCCD(T) method.

All of the calculations were performed using the MOLPRO quantum chemistry program package,[30] except the CCS-D(TQ) calculations performed with the ACES II program.[31]

Energies and Optimized Geometries of CO_3

Figure 2 shows the equipotential curves on the contour map of the APES of the ground state of CO_3 in the cross-section of the (Q_θ, Q_ε) plane of the interaction mode obtained by the BCCD(T) method (SA-CASSCF yields a similar surface, which is not shown here). Figures 3 and 4 present the energy levels of all three terms (A_1', 1E', and 2E') in the cross-section of the APES along the two interaction coordinates Q_θ and Q_ε, respectively. The CO_3 geometries in the D_{3h} and C_{2v} configurations of the APES minima are shown in Figure 5. Figure 6 shows the energy levels of all four terms (A_1', E'', 1E', and 2E') as a function of the out-of-plane a_2'' displacement. They all remain approximately harmonic, so there are no significant implications caused by the separation of the (E'' + 1E' + 2E') ⊗ a_2'' problem. The PJTE here is weak, lowering the corresponding vibrational frequencies but not producing instabilities (see the weak PJTE[17]). Similarly, the mixing of the three terms A_1', 1E', and 2E' under totally symmetric displacements, as seen from Figure 7, is also weak, but the two excited states have shifted minimum positions in accordance with eq 4 obtained below. Numerical
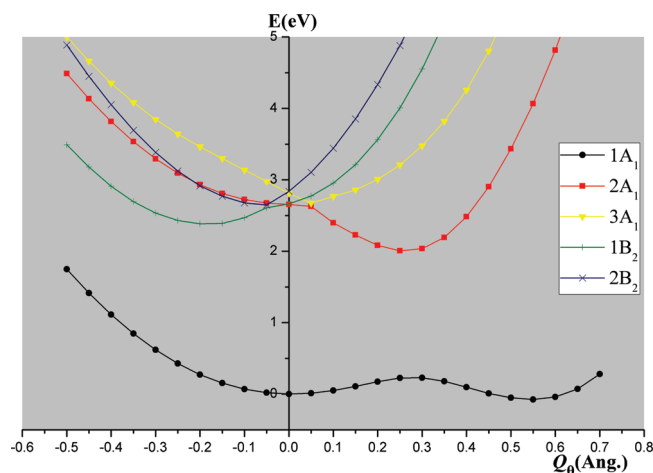
Figure 3. Ground 1A$_1$ and four excited state energy levels (2A$_1$ and 1B$_2$ and also 3A$_1$ and 2B$_2$ correspond to 1E' and 2E' in D_{3h} symmetry, respectively) of CO_3 in the cross-section of the APES along the interaction mode Q_θ obtained by the MRCI+Q//SA-CASSCF (16/13) method.
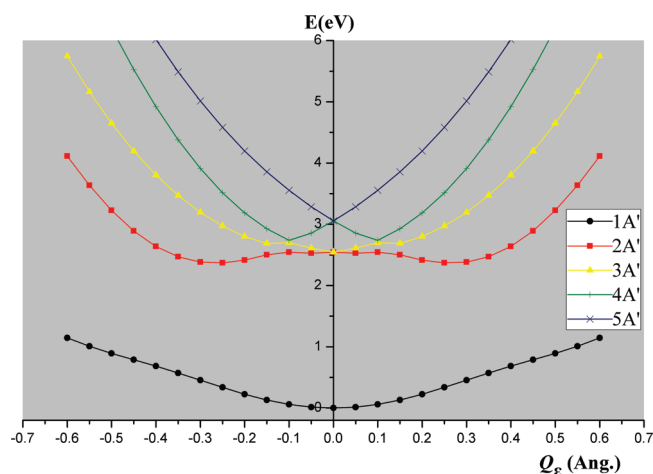
Figure 4. Ground 1A' and four excited energy levels (2A' and 3A' and also 4A' and 5A' correspond to 1E' and 2E' states in D_{3h} symmetry, respectively) of CO_3 in the cross-section of the APES along the interaction mode Q_ε obtained by the SA-CASSCF (16/13) method.

data, on geometries, energies, and frequencies of CO_3 calculated by different computational methods, are listed in Tables 1 and 2.
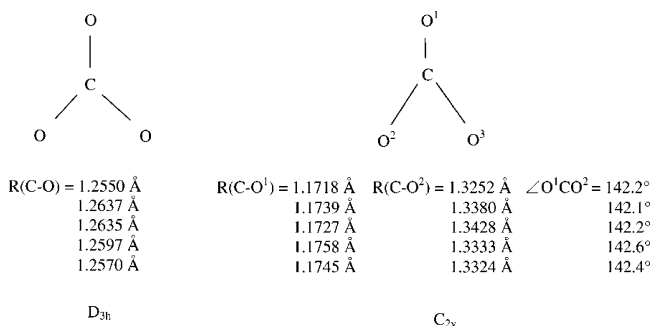
Figure 5. Optimized CO_3 geometries of the D_{3h} and C_{2v} configurations at the APES minima (bond lengths in ångstroms and angle in degrees) obtained by MRCI+Q(16/10), SS-CASSCF(16/13), SA-CASSCF(16/13), BCCD(T), and CCSD(TQ) methods, respectively.
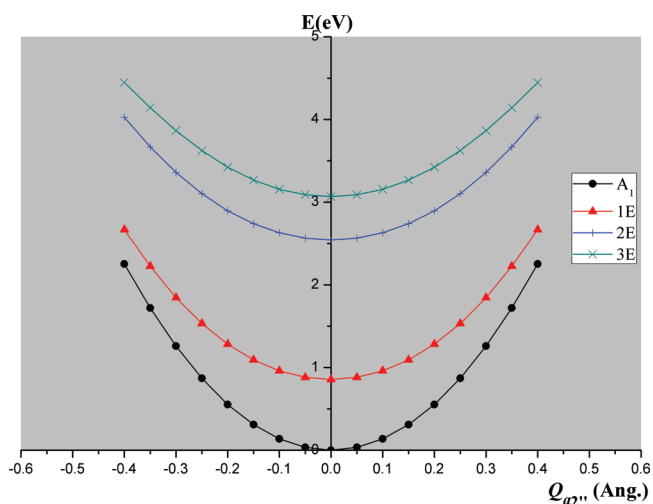
Figure 6. Energy levels of all four terms (A_1, 1E, 2E, and 3E states corresponding to the A_1', E'', 1E', and 2E' states in the D_{3h} point group) along the out-of-plane $Q_{a2''}$ displacement calculated with the SA-CASSCF(16/13) method.
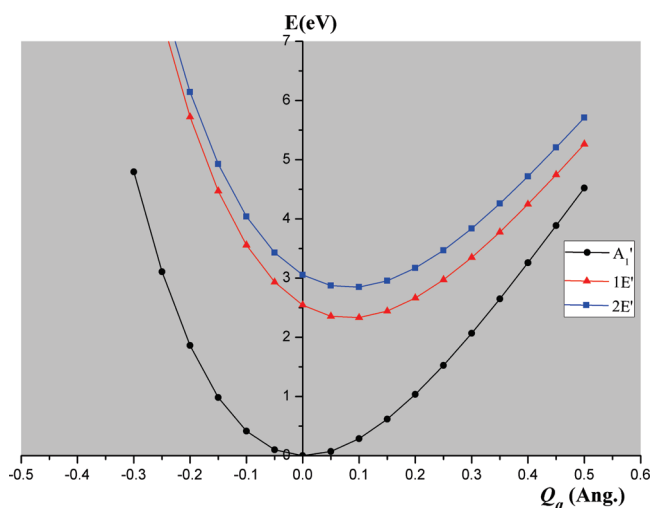
Figure 7. Energy levels of the three terms (A_1', 1E', and 2E') along the totally symmetric displacements Q_a calculated with the SA-CASSCF(16/13) method.

All of the methods used in the present study reveal two kinds of minima in the ground state APES of CO_3: the central minimum, which yields the geometrical structure of D_{3h}

symmetry, and three equivalent minima of C_{2v} symmetry. The latter have slightly lower energy than the D_{3h} minima. The single-reference methods, BCCD(T) and CCSD(TQ), produce much larger energy differences between the two minima than the multireference methods (Table 1). We also found that the results of MRCI+Q with the active space of (16/10) are similar to those of the SS-CASSCF and SA-CASSCF methods obtained with the more time-consuming (16/13) active space. It shows that, for the system under consideration, the dynamical correlation may strongly influence the relative energies but affect less the optimized geometry and frequencies.

In addition to the dominant closed-shell configuration $(4a_1')^2(1a_2'')^2(3e')^4$ $(4e')^4(1e'')^4$ in the active space with a weight of 69%, SA-CASSCF calculation shows that there is a two-electron excited configuration $(4a_1')^2(1a_2'')^2-(3e')^4(1e'')^4(4e')^2(1a_2')^2$ with a weight of 18% in the ground state, which means the single-reference method may fail. To remedy this, higher-order correlations should be included. In one of the previous papers,[16] the D_{3h} minimum was not found in the calculations by CCSD and CCSD(T) methods. Their total energies, geometries, and frequencies could not be exactly reproduced in our calculations, and the discrepancy cannot be explained by a numerical error. The results that we obtained for this case with different programs agree very well (the D_{3h} configuration was checked at different levels by ACES II,[31] MOLPRO,[30] and several other quantum chemistry programs). Arguably, this discrepancy may be caused by the defects in the HF-SCF codes of the old version of ACES-II. In our computations with the BCCD(T) and CCSD(TQ) methods, we got the D_{3h} minimum, quite similar to the results of the MRCI+Q method.

For the excited states of CO_3, SA-CASSCF (16/13) calculations also show that there is a large static correlation contribution. In particular, for the 2E' state, the dominant electronic configuration comes from a two-electron excitation of the ground state, which cannot be described well by single-reference methods, such as popular EOM-CC and TDDFT. The multireference treatment is necessary in such cases. Perhaps this is the reason that in our multirefernce calculations of the excited states the energies are lower than those obtained by other methods (e.g., in ref 16).

Further Discussion: Jahn−Teller Parameter Values

As discussed above, it is seen in Figure 1 that the Q_θ coordinate of the interaction mode coincides with the line that connects the D_{3h} and C_{2v} minima. On the other hand, we can get the coordinate of this line from the ab initio calculations following the distortions from the geometry at the D_{3h} to the C_{2v} minima in terms of the symmetrized normal displacements a_1', $1e'$, and $2e'$ in the D_{3h} geometry. This can be done by optimizing the corresponding linear combination of these coordinates to fit the numerical values of the atomic coordinates of CO_3 in the two minima. It yields (cf. eq 2) the following:

$$Q_\theta = 0.98931q_{1\theta} + 0.00479q_{2\theta} + 0.14574q_a \quad (4)$$

A Seven-State Six-Mode Problem

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2685**

Similarly,

$$Q_\varepsilon = 0.99999q_{1\varepsilon} + 0.00485q_{2\varepsilon} \tag{5}$$

The coefficients in the linear combination show that, in the distortion from the high-symmetry D_{3h} configuration to the C_{2v} one along the interaction mode, the contribution of the bend $1e'$ mode is overwhelming, and there is also a significant contribution from the totally symmetric displacements Q_a, which is understandable in view of the change in the average interatomic distances in this transformation. It is reflected also in the shift of the minimum along the Q_a coordinate in the excited E terms in Figure 7.

Comparing the computed energy levels of the five states (Figure 3) with those in Figure 1, we see that qualitatively the formation of the low-symmetry C_{2v} minima is indeed controlled by the JTE in the excited E states, similar to the O_3 case.[18] But compared to the latter, the CO_3 system has a more complicated electronic structure. First, there are three excited state conical intersections in the D_{3h} symmetry of CO_3 and a considerable PJT interaction between two of the excited states at the avoided crossing point $Q_\theta = 0.05$ Å. More important is the interaction between the E_θ branch of the excited E state with the ground state at another point of avoided crossing ($Q_\theta = 0.30$ Å) that is very strong and produces a gap of $\Delta = 1.78$ eV between them (in O_3, this interaction is very small). For this reason, the barrier between the two geometries, $\delta = 0.22$ eV, is much smaller than in O_3, but it is still significantly higher than the zero vibration level, so the two configurations can be observed spectroscopically. The large energy gap at the avoided crossing may also quench the otherwise important topological (Berry) phase implications, which influence the rovibronic spectrum in the JT minima (see, e.g., the discussion of the O_3 case[18]).

On the contour map of the ground state APES in the space $(Q_\theta, Q_\varepsilon)$ (Figure 2), the positions and the relative barriers between the central minimum and the peripheral ones and between any two C_{2v} minima are qualitatively different from many other multiminima APES in JT systems with ground state degeneracy.[17] Indeed, while in the latter cases the transition between the peripheral minima goes directly between them sidestepping (bypassing) the high-symmetry configuration, in CO_3, the direct barrier between the C_{2v} minima is very high as compared with the barrier to the D_{3h} one. This situation may be important in evaluation of the tunneling splitting, as well as the whole rovibronic spectrum.

The numerical results of the ab initio calculations allow us also to approximately estimate the values of the majority of the effective JT parameters in eq 3. To do this, we can produce further appropriate cross-sections of the multidimensional five-branched APES and compare the energies as functions of the JT parameters with ab initio obtained values. By fitting the ab initio data at small deviations from the D_{3h} symmetry with the roots of eq 3 in the cross-section $Q_\varepsilon = 0$ with $Q_a = 0$, and $Q_\theta = 0$ with $Q_\varepsilon = 0$, respectively the approximately estimated parameters are as follows: $K_1 = 18.92$ eV/Å^2, $K_2 = 30.40$ eV/Å^2, $K_3 = 14.17$ eV/Å^2, $F_{12} = 0.84$ eV/Å, $F_{13} = 2.99$ eV/Å, $F_{23} = 0.76$ eV/Å, $F_{22} = 0.73$ eV/Å, $F_{33} = 4.76$ eV/Å, $G_{22} = -4.34$ eV/Å^2, and $G_{33} = -5.70$ eV/Å^2; $K_1^a = 70.95$ eV/Å^2, $K_2^a = 80.90$ eV/Å^2, $K_3^a = 78.58$

eV/Å^2, $F_2^a = 6.50$ eV/Å, $F_3^a = 6.32$ eV/Å, and $F_{23}^a \approx 0$. The fitting is approximate and may not be acceptable for large deviations from the D_{3h} symmetry. As expected (Figure 1), the strongest PJTE comes from the interaction between the ground state and the 2E$'$ state, and the stronger JTE is in the degenerate 2E$'$ state too. Even though the interaction between the other excited 1E$'$ and the ground state A$_1'$ is small, it cannot be ignored, as it affects indirectly the ground state structure through its JT and PJT interaction with the 2E$'$ term. In total, the coexistence of the D_{3h} and C_{2v} equilibrium configurations in CO_3 is controlled by the combination of the JTE and PJTE, which mix five electronic states with five vibrational modes. The small F_i^{ae} constants of quadratic interactions between the a_1' and e' modes could not be obtained in our calculations because they fall out from the cross-sections of either $Q_a = 0$ or $Q_e = 0$, while calculation of the whole Q_aQ_e surface is extremely time-consuming and seems to be excessive.

As for the separate excited state E$'' \otimes (a_1' + 1e' + 2e')$ problem, the calculations reveal a somewhat different interaction mode along which the JTE is very weak with a distortion (the radius at the bottom of the trough of the APES[17]) of $\rho_0 \approx 0.07$ Å, which hardly influences the geometry in this state.

Conclusion

The CO_3 molecule proved to have a very complicated electronic structure and two different, coexisting geometry configurations. An electronic structure problem of this complexity with a full explanation of the origin of peculiar geometries due to the JTE and PJTE in excited states is given here for the first time. High-level ab initio calculations of the APES of the ground and six (three two-fold degenerate in the high-symmetry configuration) excited electronic states in the space of all six (two two-fold degenerate in-plane, one totally symmetric, and one out-of-plane) active vibrational modes of the CO_3 molecule were performed and rationalized in terms of a combined two-mode JT and two-mode PJT problem. One excited state plus the out-of-plane mode were separated by symmetry. The two-mode JT problem was reduced to the one-mode one by means of a coordinate transformation. The coexistence of two types of APES minima with D_{3h} symmetry (one minimum) and C_{2v} symmetry (three equivalent minima) was confirmed, and the origin of the distorted configuration as due to the strong JTE in one of the excited states was revealed. The harmonic vibrational frequencies in the minima of the two configurations were also evaluated. The efficiency of the JTE theory in rationalizing the results of ab initio calculations is demonstrated here by the solution of a very complicated problem of CO_3.

References

(1) Katakis, D.; Taube, H. *J. Chem. Phys.* **1962**, *36*, 416–422.

(2) Moll, N. G.; Clutter, D. R.; Thompson, W. E. *J. Chem. Phys.* **1966**, *45*, 4469–4481.

(3) DeMore, W. B.; Dede, C. *J. Phys. Chem.* **1970**, *74*, 2621–2625.

(4) Jacox, M. E.; Milligan, D. E. *J. Chem. Phys.* **1971**, *54*, 919–926.

(5) Yung, Y. L.; Demore, W. B.; Pinto, J. P. *Geophys. Res. Lett.* **1991**, *18*, 13–16.

(6) Yung, Y. L.; Lee, A. Y. T.; Irion, F. W.; DeMore, W. B.; Wen, J. *J. Geophys. Res.* **1997**, *102*, 10857–10866.

(7) Perri, M. J.; Van Wyngarden, A. L.; Lin, J. J.; Lee, Y. T.; Boering, K. A. *J. Phys. Chem. A* **2004**, *108*, 7995–8001.

(8) Bennett, C. J.; Jamieson, C.; Mebel, A. M.; Kaiser, R. I. *Phys. Chem. Chem. Phys.* **2004**, *6*, 735–746.

(9) Kaiser, R. I.; Mebel, A. M. *Chem. Phys. Lett.* **2008**, *465*, 1–9.

(10) Ung, A. Y.; Schiff, H. I. *Can. J. Chem.* **1966**, *44*, 1981–1991.

(11) Weissberger, E.; Breckenridge, W. H.; Taube, H. *J. Chem. Phys.* **1967**, *47*, 1764–1769.

(12) Jamieson, C. S.; Mebel, A. M.; Kaiser, R. I. *Chem. Phys. Chem.* **2006**, *7*, 2508–2513.

(13) Van de Guchte, W. J.; Zwart, J. P.; Mulder, J. J. C. *J. Mol. Struct. (THEOCHEM)* **1987**, *152*, 213–229.

(14) Castro, M. A.; Sylvio, S. *Chem. Phys. Lett.* **1991**, *177*, 98–102.

(15) Mebel, A. M.; Hayashi, M.; Kislov, V. V.; Lin, S. H. *J. Phys. Chem. A* **2004**, *108*, 7983–7994.

(16) Kowalczyk, T.; Krylov, A. I. *J. Phys. Chem. A* **2007**, *111*, 8271–8276.

(17) Bersuker, I. B. *The Jahn-Teller Effect*; Cambridge University Press: Cambridge, U. K., 2006.

(18) Garcia-Fernandez, P.; Bersuker, I. B.; Boggs, J. E. *Phys. Rev. Lett.* **2006**, *96*, 163005.

(19) Qin, C.; Soo, T.-Y. *J. Mol. Struct. (THEOCHEM)* **2009**, *897*, 32–35.

(20) Stanton, J. F. *J. Chem. Phys.* **2007**, *126*, 134309.

(21) (a) Toyozawa, Y.; Inoue, M. *J. Phys. Soc. Jpn.* **1965**, *20*, 1289–1290. (b) Toyozawa, Y.; Inoue, M. *J. Phys. Soc. Jpn.* **1966**, *21*, 1663–1679.

(22) Bersuker, I. B.; Polinger, V. Z. *Vibronic interactions in Molecules and Crystals*; Springer-Verlag: Berlin, 1989.

(23) Polinger, V. Z. Private communication, 2009.

(24) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(25) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(26) Knowles, P. J.; Werner, H.-J. *Theor. Chim. Acta* **1992**, *84*, 95–103.

(27) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053–5063.

(28) (a) Hampel, C.; Peterson, K.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1–12. (b) Deegan, M. J. O.; Knowles, P. J. *Chem. Phys. Lett.* **1994**, *227*, 321–326.

(29) (a) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513–522. (b) Raghavachari, K.; Pople, J. A.; Replogle, E. S.; Head-Gordon, M. *J. Phys. Chem.* **1990**, *94*, 5579–5586.

(30) Werner, H.-J. *MOLPRO*, version 2006.1; University College Cardiff Consultants Limited: Cardiff, Wales, 2006.

(31) Stanton, J. F. *ACES II*, version 2.5.0; University of Florida: Gainesville, FL, 2006.

JCTC Journal of Chemical Theory and Computation

Unrestricted Coupled Cluster and Brueckner Doubles Variations of W1 Theory

Ericka C. Barnes,[†] George A. Petersson,*[,†] John A. Montgomery, Jr.,[‡,¶]
Michael J. Frisch,[‡] and Jan M. L. Martin[¶,§]

*Hall-Atwater Laboratories of Chemistry, Wesleyan University, Middletown,
Connecticut 06459-0180, Gaussian, Incorporated, 340 Quinnipiac Street, Building 40,
Wallingford, Connecticut 06492-4050, and Department of Organic Chemistry,
Weizmann Institute of Science, IL-76100 Reovot, Israel*

Abstract: Unrestricted coupled cluster spin contamination corrected [UCCSD(T)] and unrestricted Brueckner doubles [UBD(T)] variations of the Weizmann-1 theory (W1), denoted as W1U, W1Usc, and W1BD, respectively, are compared with the restricted open-shell W1 theory [W1(RO)]. The performances of the four W1 variants are assessed with 220 total atomization energies, electron affinities, ionization potentials, and proton affinities in the G2/97 test set, for consistency with the error analysis of the original W1(RO) study. The root-mean-square deviations from the experiment of W1U (0.65 ± 0.48 kcal/mol), W1Usc (0.57 ± 0.48 kcal/mol), W1BD (0.62 ± 0.48 kcal/mol), and W1(RO) (0.57 ± 0.48 kcal/mol) show that the four methods are virtually indistinguishable. This error analysis excludes the "singlet biradicals," C_2 and O_3, since single determinantal methods are not really adequate for these strongly multireference systems. The unrestricted W1 variants perform poorly for such highly spin-contaminated and multireference species (the largest deviation from experiment for W1Usc is -4.2 ± 0.1 kcal/mol for the O_3 EA). W1(RO) performs much better than its unrestricted counterparts for these pathological cases (the deviation from experiment is reduced to -1.5 ± 0.1 kcal/mol for the O_3 EA), though the errors are significantly larger than those for the overall test set. The examples of C_2, O_3, and the F_2 potential energy curve indicate that an advantage to using W1BD is that the error in $\langle S^2 \rangle$ correlates with the magnitude of the error in energy, whereas W1(RO) loses accuracy without such a warning.

I. Introduction

Advances in computational methods and computer hardware have made possible the accurate ab initio calculation of energies for small- and medium-size molecules. Combined with Pople's model chemistry concept, these calculations provide reliable thermochemical predictions, which are a significant achievement of modern computational chemistry.[1]

A "theoretical model chemistry" is a complete algorithm for the calculation of the energy of any molecular system.[2,3] It cannot involve subjective decisions in its application. It must be size extensive, giving energies that are additive for separated systems, so that the energy of every molecular species is uniquely defined. A model chemistry is useful if for some class of molecules it is the most accurate calculation we can afford to do. A number of "black-box" computational methods have emerged in the past two decades, through the development of composite theoretical model chemistry methods such as the complete basis set (CBS) model chemistries of Petersson et al.,[4−8] the Gaussian-*n* methods of Pople and co-workers,[9−12] the Weizmann-*n* (W*n*) theories of Martin and co-workers,[13−17] the high-accuracy extrapo-

* Corresponding author. E-mail: gpetersson@wesleyan.edu.

[†] Wesleyan University.

[‡] Gaussian, Incorporated.

[¶] Weizmann Institute of Science

[§] Current address: Department of Physics, University of Connecticut, 2152 Hillside Road, U-3046, Storrs, CT 06269-3046

2688 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Barnes et al.

lated ab initio thermochemistry (HEAT) protocol of the Gauss and Stanton groups,[18−20] and the correlation-consistent composite approach (ccCA) of Wilson and co-workers,[21,22] to name a few.

The W*n* computational protocols of Martin and co-workers offer a sequence of models of increasing cost and accuracy,[23−26] the converging hierarchy of which currently ranges from W1 to W4.4. W1 theory, the most computationally accessible member of the evolving W*n* family, is often employed as a benchmark for more approximate methods in the absence of accurate experimental data. Some of the key accomplishments of W1 theory are:

i. It achieves 0.44 kcal/mol mean absolute deviation (0.56 kcal/mol rms deviation) for 220 total atomization energies (TAEs), electron affinities (EAs), ionization potentials (IPs), and proton affinities (PAs) of the G2/97 set;

ii. As a properly defined theoretical model chemistry, it is applicable in a "black-box" manner by a nonspecialist; and

iii. It is completely devoid of parameters adjusted to fit experimental data.

The present study compares the performance of several unrestricted variants of W1 theory: unrestricted W1 (W1U), W1U with a spin contamination correction (W1Usc), and unrestricted Brueckner doubles W1 (W1BD). Standard W1 will be explicitly referred to as W1(RO) in this study, in order to avoid confusion in the comparison.

II. Restricted vs Unrestricted Reference

One of the major issues in computational studies of open-shell species is the selection of a restricted or an unrestricted reference wave function. Each has its well-known advantages and disadvantages. For example, restricted open-shell coupled cluster singles and doubles[27−31] with perturbative triples,[32−35] [ROCCSD(T)], dissociate to the wrong energy limit, while the unrestricted coupled cluster [UCCSD(T)] wave function dissociates to the correct energy limit, but the wave function becomes significantly spin contaminated as a bond dissociates. Several approaches have been used through the years to alleviate the spin contamination problem. For example, the spin correction term in W1Usc was introduced for this purpose.[36]

Whether one chooses a restricted or an unrestricted reference determinant, it is best to be consistent. Although it is common practice in the application of unrestricted methods to treat many unrestricted Hartree−Fock (UHF)-unstable species (e.g., F_2, alkenes, or polyenes) with a restricted reference, problems arise in *reactions* involving these "closed-shell" molecules. For example, the methyl C−H bond dissociation energy (BDE) of UHF-unstable 1-butene gives a radical product, $H_2C=CHCH_2\dot{C}H_2$, that would be described by an unrestricted determinant. Employing a restricted reference for 1-butene and an unrestricted reference for the 1-buten-4-yl radical would create consistency problems between reactant and product energies and, thus, generate spurious BDE contributions. No matter how distant a radical center is, an unrestricted treatment of the radical will induce spin polarization of the UHF-unstable π-bond. These problems are eliminated if one employs a

restricted open-shell Hartree−Fock (ROHF) reference for the radical or a UHF reference for the parent alkene.

Another alternative to unrestricted Hartree–Fock (HF) and coupled cluster is the use of charge coupled device (CCD) calculations with Brueckner orbitals,[37−42] coined as "Brueckner doubles" (BD) by Handy et al.[41] This method employs a reference configuration, BDRef (from which the singles coefficients are zero for the coupled cluster wave function truncated at the doubles level), in place of the HF reference typically employed in the coupled cluster ansatz. The Brueckner condition implies that the corresponding reference determinants give the best overlap of a one-configuration approximation of the wave function with the exact wave function.[39] Such a reference was introduced by its namesake Brueckner in 1954, as part of a self-consistent method in nuclear physics,[37] and was first employed by Nesbet[38] for use in the configuration interaction (CI) expansion of a wave function. In the field of computational chemistry, the resurgence of interest in Brueckner orbitals in the past two decades was prompted by Chiles and Dykstra[40] and later by Handy and co-workers.[41] Similar studies that demonstrate the robustness of BD as an alternative to its coupled cluster counterpart[42−45] have been carried out over the years. Since the species considered in the previous paragraph (i.e., F_2 and alkenes) are not UBD-unstable, the problems of consistency between the reactant and the radical product are also eliminated by using a BD-based method.

III. Computational Details

A. Components of W1. The justification for the selection of each component calculation in W1(RO) has been presented in detail[14,15] and will not be repeated here. The W1U and W1BD methods retain the essential features of standard W1(RO) but replace the sequence of ROHF, ROCCSD, and ROCCSD(T) calculations with their spin-unrestricted counterparts for W1U for calculations involving open-shell species and with the BDRef, BD, and BD(T)[46] sequence for W1BD theory. All calculations were carried out with Gaussian 09,[47] which uses the ROCCSD(T) definition of UCCSD(T) in a basis of semicanonicalized ROHF orbitals,[35,48] with the semicanonicalization carried out before the integral transformation. The triples contribution to BD(T) is evaluated with semicanonical Brueckner orbitals.

The UB3LYP/cc-pVTZ+1d geometry optimization and frequency calculations are retained from the original W1(RO) implementation. Optimized geometries and energy components for all species are available in the accompanying "geometries.txt" and "components.pdf" files as Supporting Information. In the original implementation of W1(RO), scalar relativistic corrections were obtained with the Martin−Taylor small (MTsmall) basis sets[49,50] as one-electron Darwin and mass−velocity terms[51,52] from averaged coupled pair wave functions,[53] while corresponding components in this study were obtained with Douglas−Kroll−Hess (DKH) second-order scalar relativistic calculations[54−58] using a Gaussian nuclear model[59] (also employing the MTsmall basis sets). Spin−orbit calculations were taken directly from ref 14. The spin correction term in W1Usc,

Variations of W1 Theory

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2689**

$$\Delta E(\text{spin}) = -6.28 m E_h \times \Delta \langle S^2 \rangle_{\text{UHF}} \qquad (1)$$

minimizes the difference between ROCCSD(T) and UCCSD(T) energies for some highly spin-contaminated species.[36]

B. Implementation of Brueckner Doubles. The BD algorithm employed throughout this study involves macro-iterations to update the orbitals, wherein each macroiteration involves an integral transformation and a CCD calculation. Thus, BD calculations appear, at first glance, to be significantly more expensive than the corresponding CCSD calculations. In practice, however, W1BD calculations are only slightly more expensive than the corresponding ones in W1U or W1(RO). The two main reasons for this are as follows: i. Since W1 involves a sequence of calculations, the converged orbitals and amplitudes from each step can be used to start the next. As a result, most of the extra CC iterations required are performed in the first BD(T) calculation, which uses the smallest basis set in the sequence of calculations. For the largest basis set (i.e., the BD/aug-cc-pVQZ+2df step), typically only two or three BD macroiterations are required, and the second and later macroiterations require only a few CC iterations. ii. Calculations involving even three heavy atoms spend a significant amount of time in the (T) steps, (the only $O(N^7)$ parts of W1), and calculations on systems having four or more heavy atoms are dominated by these steps. The triples calculations have the same cost for CCSD(T) and BD(T).

The result is that the extra cost of doing W1BD over W1U or W1(RO) is in the range of 20–40% for two heavy-atom systems, about 20% for three heavy-atom systems, and going down further to less than 20% for systems with more than three heavy atoms and/or those involving second-row atoms and beyond. (W1 calculations involving two or more second-row or heavier atoms are dominated by the last calculation, which includes triples and correlates core electrons. As noted previously, the triples part of this step has the same cost for CCSD(T) and BD(T), and the BD iterations converge quickly since they use the orbitals, amplitudes, and basis set from the preceding frozen-core calculation as an initial guess.)

All but one of the correlation energy calculations in W1 use the frozen-core (FC) approximation, which substantially speeds up the calculations as compared to correlating all electrons (Full). Previous papers on the BD method have not discussed the issue of frozen-core with this model, and some programs, such as Gaussian 03, freeze the core orbitals at their initial values during BD iterations. This means that the converged BD(FC) energy varies with different initial guess orbitals. For example, in CN, the UHF solution is highly spin contaminated, while the BD(Full) reference determinant has very minimal spin contamination. Hence, a BD(FC) calculation starting from the UHF orbitals will freeze a core that is more spin polarized than that of a BD(Full) calculation and produce a different energy than, say, a BD(FC) calculation that starts with (also much less spin contaminated) B3LYP orbitals. The consistent approach to frozen-core BD calculations is to update all orbitals, including the core, during the BD iterations but to restrict the amplitudes in the CCD calculations to those involving only valence orbitals. The results are then independent of the initial orbitals, and the core is spin polarized only to the extent that the BD valence is. This approach is used in all BD results reported here.[47]

IV. Results and Discussion

Since the four flavors of W1 theory are virtually the same for closed-shell species, we shall focus on their performance for open-shell systems. The extent to which the ROHF and UHF reference determinants differ can be measured by the error in $\langle S^2 \rangle$ for the reference configuration.

A. Spin-Contaminated Species. The restricted and unrestricted variations of W1 theory have been evaluated with bond dissociation enthalpies (BDEs), EAs, and IPs of some radicals and biradicals (Table 1). The sample of 11 reactions was selected on the basis of sizable errors in $\langle S^2 \rangle_{\text{UHF}}$ and the availability of reliable experimental data.[60-67]

The rms deviation from experiment (Table 1) is not surprisingly the largest for W1U (1.3 ± 0.3 kcal/mol), in the absence of any rectification of the spin contamination problem. The spin correction in W1Usc indeed reduces the rms error to 0.6 ± 0.3 kcal/mol but is not always reliable. For example, $\Delta E(\text{spin})$ in eq 1 overestimates the correction for the first C–H bond dissociation in acetylene by 0.8 ± 0.02 kcal/mol and worsens the deviation with experiment of the first C–H bond dissociation of propene by 0.8 ± 0.4 kcal/mol (Table 1). The rms deviations for W1U, W1Usc, W1BD, and W1(RO) are 1.3 ± 0.3, 0.6 ± 0.3, 0.8 ± 0.3, and 0.6 ± 0.3 kcal/mol, respectively, demonstrating the comparable accuracy of W1Usc, W1BD, and W1(RO) in handling these UHF spin-contaminated species. Note that both W1BD and W1(RO) attain these results in the *absence* of the empirical spin correction in W1Usc.

The following interesting observations are made for the three most spin-contaminated species of the G2/97 test sets, C_2, O_3, and CS^+, selected on the basis of $\Delta \langle S^2 \rangle_{\text{UHF}}$ values greater than 0.6:[68]

i. First is the existence of multiple solutions, which complicates the use of "black-box" methods. To compound the problem of multiple solutions for both restricted and unrestricted versions of both HF and Brueckner determinants for C_2, the relative energies of the solutions are switched between the HF vs the CCSD(T) levels of theory for both restricted and unrestricted reference configurations. That a single reference method produces several solutions comes as no surprise in these cases, since C_2 and O_3 are known to possess a significant multiconfigurational character in their wave functions. If there are indications of near degeneracies of configurations (such as a large error in $\langle S^2 \rangle$ for UHF or UBDRef or a negative eigenvalue for a virtual orbital as in neutral C_2), then it is necessary to explore multiple solutions in addition to checking for the correct number of imaginary frequencies. A more thorough discussion is presented as Supporting Information (see "Problem_species.pdf" file), which includes an example of solutions crossing between the BD vs BD(T) levels of theory.

ii. Second, the reactions involving these severely spin-contaminated systems indicate that $\Delta \langle S^2 \rangle_{\text{UHF}}$ values do not correlate as well with the energy error in a W1U calculation

Table 1. Deviations (Experiment−theory) of Calculated BDEs and adiabatic IPs and EAs from Experiment of Selected Spin-Contaminated Species, in kcal/mol

		W1U $\Delta\langle S^2\rangle$		W1BD $\Delta\langle S^2\rangle$		deviation (experiment−theory)				experiment
species[a]	reaction	reactant	product	reactant	product	W1U	W1Usc	W1BD	W1RO[b]	
					σ-Radicals					
BDE ΔH_{298}	H−CN → H + ·C≡N	0.00	0.37	0.00	0.01	−1.9	−0.4	−1.0	−0.8	126.29 ± 0.2[c]
	H−C≡CH → H + ·C≡CH	0.00	0.35	0.00	0.03	−0.6	0.8	0.0	0.1	133.46 ± 0.02[d]
	H−CH=CH₂ → H + ·CH=CH₂	0.00	0.18	0.00	0.01	−0.4	0.3	−0.1	−0.1	110.71 ± 0.6[c]
IP ΔE_0	C=O → e⁻ + C=O⁺	0.00	0.21	0.00	0.02	−0.9	−0.1	−0.6	−0.2	323.17 ± 0.01[e]
	N≡C−C≡N → e⁻ + N≡C−C≡N⁺	0.00	0.35	0.00	0.03	−1.8	−0.4	−1.3	−1.0	308.42 ± 0.2[f]
EA ΔE_0	CH₂=C=CH⁻ → e⁻ + CH₂=C=CH	0.00	0.21	0.00	0.03	−0.7	0.1	−0.4	−0.3	20.59 ± 0.1[f]
	C=N → e⁻ + C=N	0.00	0.37	0.00	0.01	−1.8	−0.3	−1.0	−0.7	89.06 ± 0.1[e]
					π-Radicals					
BDE ΔH_{298}	HCH₂−CH=CH₂ → H + ·CH₂−CH=CH₂	0.00	0.20	0.00	0.04	0.5	1.3	0.8	0.9	88.79 ± 0.4[g]
EA ΔE_0	CH₂CHCH₂⁻ → e⁻ + CH₂CHCH₂	0.00	0.20	0.00	0.04	−0.5	0.3	−0.3	−0.1	11.09 ± 0.2[f]
					Singlet Biradicals					
BDE ΔH_{298}	F−F → F + F	0.30	0.01	0.00	0.01	1.8	0.6	1.1	1.0	38.00 ± 0.2[f]
					Triplet Biradicals					
EA ΔE_0	H−C−C≡N⁻ → e⁻ + H−C−C≡N	0.12	0.36	0.02	0.07	−0.5	0.5	−0.3	−0.2	46.20 ± 0.3[h]
MAD						1.1	0.5	0.6	0.5	± 0.2
rms						1.3	0.6	0.8	0.6	± 0.3
LD						−2.1	1.3	−1.3	−1.0	

[a] Geometries are optimized at the UB3LYP/cc-pVTZ+1d level and available as Supporting Information. [b] W1(RO) energies were calculated with Gaussian 09 (ref 47). [c] ref 60. [d] ref 61. [e] ref 62. [f] ref 63. [g] ref 64. [h] ref 65.

as $\Delta\langle S^2\rangle_{UBDRef}$ values do for W1BD errors (vide infra). A sizable $\Delta\langle S^2\rangle_{UBDRef}$ value appears to be a useful warning of when to be skeptical of W1BD results (refer to Supporting Information Table S−III and discussion on pages 11−12 of "Problem_species.pdf" file for details).

B. G2/97 Test Set. The G2−1[9−11] and the G2−2[69,70] test sets, collectively referred to as the G2/97 data set of Curtiss et al. were employed to calibrate the accuracy of W1(RO) for EAs, IPs, and PAs, while the G2−1 test set and a subset of the G2−2 data set (26 out of 93) of heats of formation were used in the calibration of total atomization energies (TAEs).[14] The selection of TAEs, EAs, IPs, and PAs is retained in assessing the performance of W1U, W1Usc, and W1BD, in order to facilitate comparisons with W1(RO) values in the literature.[14,15] A thorough discussion on the notable discrepancies with experiment for some problematic cases has already been presented by one of the authors.[14] A comprehensive breakdown of the G2−1 and the G2−2 TAE, EA, IP, and PA energetic components and error statistics for W1U, W1Usc, W1BD, and W1(RO) are available as Supporting Information. The error analyses for 220 reactions from the G2/97 data set are summarized in Table 2. The reported uncertainties in the deviations from experiment represent the uncertainties in the experimental data.[36]

The overall G2/97 rms errors for W1U, W1Usc, W1BD, and W1(RO) (excluding C₂ and O₃) are 0.65, 0.57, 0.62, and 0.57 ± 0.48 kcal/mol, respectively. Comparison of the calculated energies with experiment is rather problematic due to the large experimental uncertainties (± 0.65 for TAEs, ± 0.32 for EAs, and ± 0.35 kcal/mol for IPs, Table 2). The experimental uncertainties are of greater magnitude than the differences in rms errors between the methods themselves, making W1 variants virtually indistinguishable from one another for the G2/97 test set (Table 2).

C. F₂ Potential Energy Curve. Although the W*n* methods are intended for the thermochemistry of molecular systems at equilibrium geometries, the potential energy curve for the F₂ $^1\Sigma_g^+$ ground-state dissociation provides insight into the differences between the UCCSD(T), the spin-corrected UCCSD(T), the UBD(T), and the ROCCSD(T) levels of theory. Deviations of the potential energy curves from the exact solution in the cc-pVDZ basis are shown in Figure 1. A constant geometry-independent shift of the energy from the full configuration interaction (FCI) reference would display as zero error throughout the potential energy curve.

Neither restricted nor unrestricted CCSD(T) give a reliable description of bond-breaking reactions. These qualitative features of restricted and unrestricted methods along a potential energy curve for bond dissociation are well-known.[71] The spin correction successfully reduces the UCCSD(T) error in energy but shows fluctuation and dissociates improperly, reminiscent of problems encountered with a restricted reference (Figure 1). Furthermore, Ochterski et al. pointed out that the spin contamination error increases linearly with $\Delta\langle S^2\rangle_{UHF}$ for $\Delta\langle S^2\rangle_{UHF}$ less than 0.6.[68] This is consistent with our observations for C₂, O₃, and CS⁺ (see "Problem_species.pdf" file given as Supporting Information) and strongly suggests that eq 1 is also not applicable for $\Delta\langle S^2\rangle_{UHF} > 0.6$ on the F₂ potential energy curve ($\Delta\langle S^2\rangle_{UHF} = 0.6$ is marked by the vertical dashed line at 1.5 Å in Figure 1).

The Brueckner doubles approach possesses compensating advantages over both the RHF- and the UHF-based methods. First, there is very good agreement between the UBD(T) and the ROCCSD(T) potential energy curves in the vicinity of the equilibrium geometry, where UCCSD(T) is already contaminated with unwanted contributions from higher spin multiplicities. This efficacy of Brueckner orbitals in reducing spin contamination over a wider range of geometries compared to UHF-based methods is also well documented.[72] Furthermore, the Brueckner doubles curve is practically indistinguishable from that of ROCCSD(T), up to the point where the UBD(T) energy errors exhibit extreme sensitivity

Variations of W1 Theory

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2691**

Table 2. Error Analysis (kcal/mol) for TAEs (298 K), EAs, IPs, and PAs of Select Molecules in the G2/97 Test Set for the Different Variations of W1 Theory[a]

G2/97 subset[b]	method	MAD	rms	LD	species/LD
TAE$_{298K}$ (81)[c]					
	W1U	0.60 ± 0.41	0.75 ± 0.65	1.72 ± 0.10	ClNO
	W1Usc	0.56 ± 0.41	0.70 ± 0.65	1.79 ± 0.10	ClNO
	W1BD	0.61 ± 0.41	0.77 ± 0.65	1.97 ± 0.10	ClNO
	W1(RO)	0.55 ± 0.41	0.67 ± 0.65	1.75 ± 0.10	ClNO
EA (55)[d]					
	W1U	0.46 ± 0.22	0.60 ± 0.32	−1.96 ± 0.60	CH$_2$NC
	W1Usc	0.41 ± 0.22	0.53 ± 0.32	−1.64 ± 0.60	CH$_2$NC
	W1BD	0.43 ± 0.22	0.55 ± 0.32	−1.81 ± 0.60	CH$_2$NC
	W1(RO)	0.43 ± 0.22	0.53 ± 0.32	−1.74 ± 0.60	CH$_2$NC
IP (76)[e,f]					
	W1U	0.41 ± 0.16	0.57 ± 0.35	−2.10 ± 0.23	CS
	W1Usc	0.33 ± 0.16	0.42 ± 0.35	1.39 ± 0.05	P$_2$
	W1BD	0.37 ± 0.16	0.49 ± 0.35	−1.48 ± 0.18	N$_2$ ($^2\Sigma$ cation)
	W1(RO)	0.35 ± 0.16	0.46 ± 0.35	1.42 ± 0.05	P$_2$
PA (8)[g]					
	W1U	0.42	0.48	−0.83	C$_2$H$_2$
	W1BD	0.42	0.49	−0.88	C$_2$H$_2$
	W1(RO)	0.43	0.49	−0.83	C$_2$H$_2$
Total (220)					
	W1U	0.49 ± 0.27	0.65 ± 0.48		
	W1Usc	0.44 ± 0.27	0.57 ± 0.48		
	W1BD	0.48 ± 0.27	0.62 ± 0.48		
	W1(RO)	0.44 ± 0.27	0.57 ± 0.48		

[a] C$_2$ and O$_3$ are excluded in all statistics. Optimized geometries, total energies, and energy changes of species and reactions in the G2/97 sets are given in the Supporting Information "geometries.txt", "components.pdf", and "Test_set.pdf" files, respectively. The uncertainties in the deviations from experiment represent the uncertainties in the experimental data (see "Test_set.pdf"). [b] The number of reactions considered in each subset is indicated as the number in the parentheses in the first column. [c] See Supporting Information Tables S−III and S−IV for individual TAEs. [d] See Supporting Information Tables S−V and S−VI for individual EAs. [e] See Supporting Information Tables S−VII and S−VIII for individual IPs. [f] Excludes CN ($^3\Pi$), CN ($^1\Sigma^+$), B$_2$H$_4$, sec−C$_3$H$_7$, and Si2H$_6$ (see ref 14). [g] See Supporting Information Table S–IX for individual PAs. Calculated proton affinities are considered converged at the W1 level (ref 14), agreeing well with experiment.
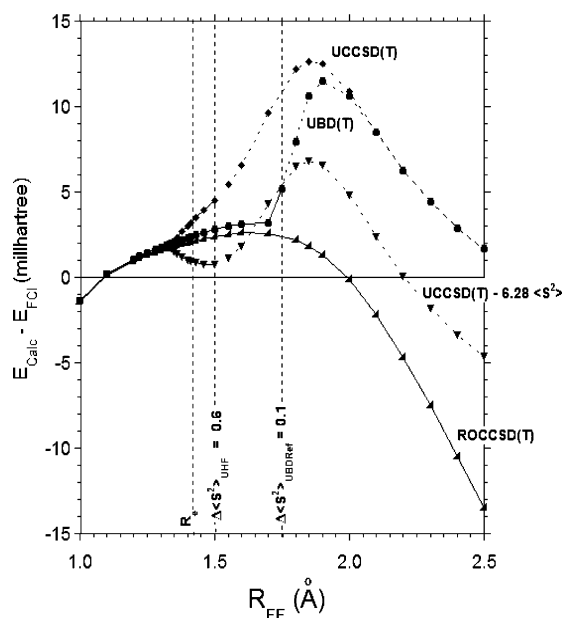
Figure 1. The error (mE_h) in the UCCSD(T), the spin-corrected UCCSD(T), the UBD(T), and the ROCCSD(T)/cc-pVDZ potential energy curves for the F$_2$ $^1\Sigma_g^+$ ground state. W1U/W1Usc and W1BD results beyond $\Delta\langle S^2\rangle_{UHF}$ = 0.6 (1.5 Å) and $\Delta\langle S^2\rangle_{UBDRef}$ = 0.1 (1.75 Å), respectively, (indicated by dashed rather than solid curves) are considered unreliable. Spin contamination errors cease to have a linear relationship with energetic errors beyond these cut-offs.

to geometry changes and the restricted coupled cluster begins to break down (1.2R$_e$ or 1.75 Å in Figure 1). This RBD(T)/UBD(T) instability point is an unambiguous demarcation between the absence and the onset of significant BD energy errors, holding promise as a diagnostic tool for assessing the reliability of BD calculations. The good correlation between the W1BD $\Delta\langle S^2\rangle$ values and the energy errors is consistent with the results for CS$^+$ and O$_3$. The potential energy curve for the restricted wave function proceeds toward the wrong energy limit without such a warning.

V. Conclusions

The unrestricted flavors of W1 theory presented in this study are viable alternatives to W1(RO). W1U benefits significantly from a spin correction term for the 11 spin-contaminated reactions (Table 1), reducing the rms error from 1.3 to 0.6 ± 0.4 kcal/mol. The accuracy of the W1Usc, the W1BD, and the W1(RO) theories are indistinguishable from one another when evaluated with the moderately spin-contaminated data set (Table 1) and the 220 total atomization energies, electron affinities, ionization potentials, and proton affinities in the G2/97 test set (Table 2). Unlike W1U, the W1BD method demonstrates good correlation between the $\Delta\langle S^2\rangle_{UBDRef}$ and the energetic errors and, thus, gives a clear indication of the onset of energetic errors associated with spin contamination. W1(RO) gives no such straightforward warning when its single determinant reference wave function is inadequate. We, therefore, recommend the use

of W1BD, rather than W1Usc, as an unrestricted alternative to W1(RO) in studies of potentially pathological cases. Multiple solutions were obtained for both the restricted and the unrestricted methods for C_2 and O_3, demonstrating the need for exploring multiple solutions with single-reference wave functions even in the "black-box" context. Of course, testing for the correct number of imaginary vibrational frequencies is always necessary.

Supporting Information Available: Optimized geometries, energy components, and chemical energy differences for all species in this study, as well as additional discussion on the problematic species C_2, O_3, and CS^+, are available in the accompanying "geometries.txt", "components.pdf", "Test-_set.pdf", and "Problem_species.pdf" files, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) For recent reviews on accurate computational thermochemistry, see for example:Helgaker, T.; Klopper, W.; Halkier, A.; Bak, K. L.; Jørgensen, P. Olsen, J. In *Quantum-Mechanical Prediction of Thermochemical Data*, Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp. 1–30; Martin, J. M. L.; Parthiban, S. In *Quantum-Mechanical Prediction of Thermochemical Data*, Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp. 31–65; Raghavachari, K.; Curtiss, L. A. In *Quantum-Mechanical Prediction of Thermochemical Data*, Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp. 67–98; Petersson, G. A. In *Quantum-Mechanical Prediction of Thermochemical Data*, Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp. 99–130; Henry, D. J.; Radom, L. In *Quantum-Mechanical Prediction of Thermochemical Data*, Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp. 161–197.

(2) Pople, J. A. Theoretical Models for Chemistry. *Proceedings of the Summer Research Conference on Theoretical Chemistry, Energy Structure and Reactivity*, 1973; p 51.

(3) (a) Pople, J. A.; Binkley, J. S.; Seeger, R. *Int. J. Quantum Chem. Symp.* **1976**, *10*, 1–19. (b) Krishnan, R.; Pople, J. A. *Int. J. Quantum Chem.* **1978**, *14*, 91–100.

(4) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. *J. Chem. Phys.* **1988**, *89*, 2193–2218.

(5) Montgomery, J. A., Jr.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1994**, *101*, 5900–5909.

(6) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A., Jr. *J. Chem. Phys.* **1996**, *104*, 2598–2619.

(7) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822–2827.

(8) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532–6542.

(9) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622–5629.

(10) Curtiss, L. A.; Jones, C.; Trucks, G. W.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1990**, *93*, 2537–2545.

(11) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221–7230.

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

(13) Martin, J. M. L.; Oliveira, G. D. *J. Chem. Phys.* **1999**, *111*, 1843–1856.

(14) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014–6029.

(15) Martin, J. M. L.; Parthiban, S. In *W1 and W2 theory and their variants: thermochemistry in the kJ/mol accuracy range*; Cioslowski, J., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2001; pp 1–30.

(16) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129–4141.

(17) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108−1144108−17.

(18) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599–11613.

(19) Bomble, Y. J.; Vázquez, J.; Kállay, M.; Michauk, C.; Szalay, P. G.; Császár, A. G.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2006**, *125*, 064108−1−064108−8.

(20) Harding, M. E.; Vázquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2008**, *128*, 114111−1114111−15.

(21) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *124*, 11410−1114104−17.

(22) DeYonker, N. J.; Grimes, T.; Yockel, S.; Dinescu, A.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *125*, 104111−1104111−15.

(23) Karton, A.; Ruscic, B.; Martin, J. M. L. *J. Mol. Struct. (THEOCHEM)* **2007**, *811*, 345–353.

(24) Karton, A.; Taylor, P. R.; Martin, J. M. L. *J. Chem. Phys.* **2007**, *127*, 064104−1−064104−11.

(25) Karton, A.; Martin, J. M. L. *J. Phys. Chem. A* **2007**, *111*, 5936–5944.

(26) Karton, A.; Martin, J. M. L. *Mol. Phys.* **2007**, *105*, 2499–2505.

(27) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

(28) Bartlett, R. J.; Purvis, G. D. *Int. J. Quantum Chem.* **1978**, *14*, 561–581.

(29) Rittby, M.; Bartlett, R. *J. Phys. Chem.* **1988**, *92*, 3033–3036.

(30) Lauderdale, W. J.; Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J. *Chem. Phys. Lett.* **1991**, *187*, 21–28.

(31) Bartlett, R. J. In *Coupled-Cluster Theory: An Overview of Recent Developments*; Yarkony, D. R., Ed.; World Scientific, Singapore, 1995; pp 1047−1131.

(32) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

(33) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

(34) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

Variations of W1 Theory

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2693**

(35) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718–8733.

(36) Wood, G. P. F.; Radom, L.; Petersson, G. A.; Barnes, E. C.; Frisch, M. J.; Montgomery, J. A., Jr. *J. Chem. Phys.* **2006**, *125*, 094106−1−094106−16.

(37) Brueckner, K. A. *Phys. Rev.* **1954**, *96*, 508–516.

(38) Nesbet, R. K. *Phys. Rev.* **1958**, *109*, 1632–1638.

(39) Larsson, S.; Smith, V. *Phys. Rev.* **1969**, *178*, 137–152.

(40) Chiles, R. A.; Dykstra, C. E. *J. Chem. Phys.* **1981**, *74*, 4544–4556.

(41) Handy, N. C.; Pople, J. A.; Head-Gordon, M.; Raghavachari, K.; Trucks, G. W. *Chem. Phys. Lett.* **1989**, *164*, 185–192.

(42) Stanton, J. F.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1992**, *97*, 5554–5559.

(43) Hampel, C.; Peterson, K. A.; Werner, H. J. *Chem. Phys. Lett.* **1992**, *190*, 1–12.

(44) Barnes, L. A.; Lindh, R. *Chem. Phys. Lett.* **1994**, *223*, 207–214.

(45) Watts, J. D.; Bartlett, R. J. *Int. J. Quantum Chem.* **1994**, *52*, 195–203.

(46) Scuseria, G. E. *Chem. Phys. Lett.* **1994**, *226*, 251–256.

(47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M. Knox, J. E. Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision A.2,* Gaussian, Inc.: Wallingford, CT, 2009.

(48) Knowles, P. J.; Andrews, J. S.; Amos, R. D.; Handy, N. C.; Pople, J. A. *Chem. Phys. Lett.* **1991**, *186*, 130–136.

(49) Martin, J. M. L.; Taylor, P. R. *Chem. Phys. Lett.* **1994**, *225*, 473–479.

(50) Martin, J. M. L. *Chem. Phys. Lett.* **1995**, *242*, 343–350.

(51) Cowan, R. D.; Griffin, D. C. *J. Opt. Soc. Am.* **1976**, *66*, 1010–1014.

(52) Martin, R. L. *J. Phys. Chem.* **1983**, *87*, 750–754.

(53) Gdanitz, R. J.; Ahlrichs, R. *Chem. Phys. Lett.* **1988**, *143*, 413–420.

(54) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89–155.

(55) Hess, B. A. *Phys. Rev. A* **1985**, *32*, 756–763.

(56) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742–3748.

(57) Barysz, M.; Sadlej, A. J. *J. Mol. Struct. (THEOCHEM)* **2001**, *573*, 181–200.

(58) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48–53.

(59) Visscher, L.; Dyall, K. G. *At. Data and Nucl. Data Tables* **1997**, *67*, 207–224.

(60) Ervin, K. M.; DeTuri, V. F. *J. Phys. Chem. A* **2002**, *106*, 9947–9956.

(61) Mordaunt, D. H.; Ashfold, M. N. R. *J. Chem. Phys.* **1994**, *101*, 2630–2631.

(62) Afeefy, H. Y.; Liebman, J. F.; Stein, S. E. NIST Chemistry Webbook. In *NIST Standard Reference Database*; Mallard, W. G., Linstrom, P. J., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, 1998; Vol. 69.

(63) *CODATA Key Values for Thermodynamics*; Cox, J. D., Wagman, D. D., Medvedev, V. A., Eds.; Hemisphere Pub. Corp: New York, 1989.

(64) Ellison, G. B.; Davico, G. E.; Bierbaum, V. M.; DePuy, C. H. *Int. J. Mass Spectrom. Ion Process.* **1996**, *156*, 109–131.

(65) Nimlos, M. R.; Davico, G.; Geise, C. M.; Wenthold, P. G.; Lineberger, W. C.; Blanksby, S. R.; Hadad, C. M.; Petersson, G. A.; Ellison, G. B. *J. Chem. Phys.* **2002**, *117*, 4323–4339.

(66) Blanksby, S. J.; Ellison, G. B. *Acc. Chem. Res.* **2003**, *36*, 255–263.

(67) Wenthold, P. G.; Squires, R. R. *J. Am. Chem. Soc.* **1994**, *116*, 6401–6412.

(68) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A., Jr. *J. Chem. Phys.* **1995**, *104*, 2598–2619.

(69) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(70) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 42–55.

(71) Schlegel, H. B. *J. Chem. Phys.* **1986**, *84*, 4530–4534.

(72) See, for example Watts, J. D.; Bartlett, R. J. *Int. J. Quantum Chem.* **1994**, *52*, 195–203.

(73) Catlett, C. In *TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications*; Grandinetti, L., Ed.; Advances in Parallel Computing Series; IOS Press: Amsterdam, The Netherlands, 2007.

JCTC Journal of Chemical Theory and Computation

Determination of London Susceptibilities and Ring Current Intensities using Conjugated Circuits

Marcos Mandado*

Department of Physical Chemistry, University of Vigo, As Lagoas (Marcosende), 36310 Vigo, Galicia, Spain

Received June 4, 2009

Abstract: Conjugated circuits have been employed to accurately reproduce the aromatic stabilization, London susceptibility, and ring current maps in polycyclic aromatic hydrocarbons, PAHs, focusing on polybenzenoids. Starting from a wave function ansatz, constructed as a superposition of Kekulé valence bond structures, the conjugated circuit resonance energy is derived using the second quantization formulation. Approximated expressions for the resonance energy, London susceptibility, and ring electron current intensity have been obtained. In these expressions, the benzene molecule is employed as a reference for the calculation in a graph theoretical fashion of properties in larger polybenzenoids. Comparison of the results obtained with conjugated circuits with those obtained using more accurate quantum chemical methods reflects the power of the conjugated circuit concept as a quantitative tool for the study of magnetic properties in PAHs. Besides the usefulness of this methodology for understanding and interpreting both the aromatic stabilization and the magnetic behavior of small and medium size PAHs, it provides a straightforward alternative way for the computation of these properties in giant PAHs for which ab initio calculations are not applicable.

1. Introduction

As defined by Randić in its essential review[1] on the application of the chemical graph theory to the study of polycyclic aromatic hydrocarbons, PAHs, conjugated circuits are those circuits within an individual Kekulé valence structure in which there is a regular alternation of both CC double and single bonds. This simple concept has been employed independently by Randić[2−4] and Gomes and Mallion[5] to develop the conjugated circuits model, CCM. The CCM performs a partitioning of the resonance energy, RE, of a PAH into local ring terms. Therefore, it allows interpreting the total aromatic stabilization of a given PAH in terms of ring contributions.[1]

A quantum chemical justification of the CCM is obtained from the valence bond theory, if one restricts the large number of valence structures present in a polycyclic conjugated hydrocarbon to the Kékule structures. This idea was first coined by Simpson[6] and subsequently employed by Herndon[7,8] more than 30 years ago to calculate the resonance

energies in PAHs from empirical quantities. On the other hand, Klein and Trinajstić[9] outlined a quantum chemical derivation of the conjugated circuits model through the Pauling−Wheland resonance theory.[10] As remarked by Randić,[1] this connection between the CCM and the quantum chemical theory is crucial because it has provided the model with a more quantitative character and a theoretical foundation.

In spite of the huge number of published articles employing conjugated circuits,[11] its scope as a quantitative chemical tool for studying the properties of PAHs is still to be exploited in its entirety. For instance, although some qualitative relations between conjugated circuits and ring current maps were pointed out some time ago by Randić,[12] the conjugated circuits have never been employed, neither for calculating magnetic susceptibilities nor for constructing ring current maps.

Alternatively other graph theoretical approaches have been developed to determine the London susceptibility,[13] i.e., the magnetic susceptibility due to induced ring electron currents. The efforts of Aihara in this direction deserve special mention. He was able to connect the aromatic stabilization

---

* Corresponding author. E-mail: mandado@uvigo.es.
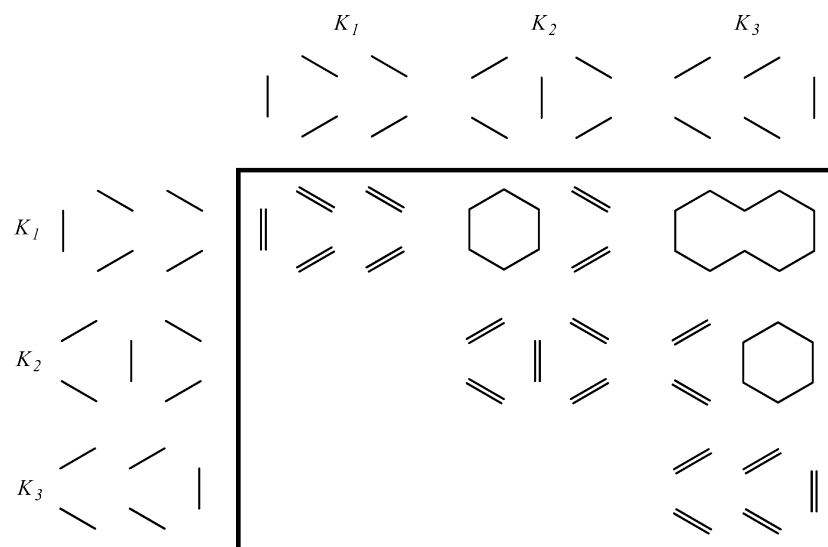
Figure 1. Superposition diagrams for the Kekulé structures of naphthalene.

with the diamagnetic susceptibility exaltation[14−17] and the induced ring electron current[18] in polycyclic conjugated systems through the extended Hückel−London theory of diamagnetism and the definition of the topological resonance energy.[19,20] His works first demonstrated that the magnetic and energetic criteria of aromaticity are indeed interrelated,[21,22] showing that the London susceptibility of a PAH is the result of the superposition of individual ring circuit contributions. However, the application of the Aihara's methodology to compute individual and total London susceptibilities requires the calculation of the circuit resonance energies,[23] which is restricted to the Hückel molecular orbital theory[24,25] and can become unmanageable for the case of giant PAHs.

On the contrary, because of its simplicity and intuitiveness, the conjugated circuits model can be considered the most powerful graph theoretical tool for the study of the aromatic stabilization in PAHs. In addition, it can provide more accurate results, with no limit for the system size, than those obtained by other topological approaches, since it can be applied in combination with more accurate quantum chemical methods. In this paper, we extend the application of conjugated circuits to the calculation of London susceptibilities and ring current intensities. In the following section, Section 2, the conjugated circuit resonance energy is derived using an operator formalism developed in the context of the second quantization, in order to lend the necessary theoretical foundation to the method. Then, the expressions for the London susceptibility and the ring current intensity are obtained. Comparison of the results with those obtained using more accurate quantum chemical methods is performed in Section 3. Finally, the main conclusions are formulated in the last section, Section 4.

2. Background and Theoretical Development

Conjugated Circuit Resonance Energy. In this section, the conjugated circuit resonance energy is obtained using a formulation based on creation and annihilation operators. Using this formulation, the Kekulé valence bond structures and a series of operators associated to the conjugated circuits

can be represented explicitly. Our derivation starts with the same wave function ansatz as that proposed by Herndon,[8] where the normalized ground-state wave function (eq 1) is expressed by a linear combination of equally weighted normalized functions, each corresponding to a Kekulé structure:

$$|\Psi\rangle = \frac{1}{\sqrt{n_K}} \sum_{i=1}^{n_K} |K_i\rangle \tag{1}$$

where n_K is the total number of Kekulé structures. The expectation value of the energy for a given PAH is then obtained from eq 2:

$$E = \frac{1}{n_K} \sum_{i=1}^{n_K} \sum_{j=1}^{n_K} \langle K_i | \hat{H} | K_j \rangle \tag{2}$$

The different contributions to the energy that are implicit in eq 2 can be viewed straightforwardly using the superposition diagrams of Kekulé structures. The superposition diagrams for the Kekulé structures of naphthalene are shown in Figure 1. As can be seen, the superposition of a Kekulé structure with itself results in a set of localized bonds, whereas the superposition of two different Kékule structures results in a sum of localized bonds and cyclic delocalized structures. These cyclic delocalized structures result from the superposition of a conjugated circuit L with its counterpart L', which only differ in the arranging of the alternant bonds. One can represent the energy contribution of a cyclic delocalized structure by introducing appropriate operators into eq 2. In order to shape such operators, it is useful to employ the second quantization formulation of the valence bond theory.[26] So, the kets $|K_i\rangle$ in eq 1 can be represented in second quantization by

$$|\Psi\rangle = \frac{1}{\sqrt{n_K}} \sum_{i=1}^{n_K} \hat{K}_i^+ |0\rangle \tag{3}$$

2696 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Mandado

$$\hat{K}_i^+ = \prod_{n=1}^{N_{e/2}} \Delta_{KL(i,n)}^{\alpha\beta+} \tag{4}$$

where each operator $\Delta_{KL}^{\alpha\beta+}$ acts on the vacuum state $|0\rangle$ by creating an electron pair in the bond formed by atomic orbitals K and L; this operator is in turn an antisymmetrized product of two particle creation operators:[26]

$$\Delta_{KL}^{\alpha\beta+} = a_{K\alpha}^+ a_{L\beta}^+ - a_{K\beta}^+ a_{L\alpha}^+ \tag{5}$$

where the subindex α or β indicates the electron has spin alpha or beta, respectively. The set of KL atomic orbitals necessary to construct the operator $\hat{K}_i^+$ depends on each Kekulé structure, but its number is always equal to $N_e/2$, N_e being the number of π electrons. The nomenclature introduced in eq 4 to denote each pair of atomic orbitals KL employs two subindices, i to indicate that the bond is associated to the Kékule structure i, and n to distinguish between different bonds associated to the same Kekulé structure, so that two sets of different subindices (i,n) may refer to the same pair of atomic orbitals provided that these subindices are defined for different Kekulé structures.

The application of the set of $\hat{K}_i^+$ operators of a given PAH on the vacuum state $|0\rangle$ generates the kets $|K_i\rangle$ that represents the Kekulé structures of the system. Therefore, they will be called Kekulé operators from now on. In addition, the bra $\langle K_i|$ can be generated through the adjoint of $\hat{K}_i^+$, $\hat{K}_i$,

$$\langle \Psi | = \frac{1}{\sqrt{n_K}} \sum_{i=1}^{n_K} \langle 0 | \hat{K}_i \tag{5a}$$

$$\hat{K}_i = \prod_{n=1}^{N_{e/2}} \Delta_{KL(i,n)}^{\alpha\beta} \tag{6}$$

where each operator $\Delta_{KL}^{\alpha\beta}$ acts on the vacuum state $\langle 0|$ by creating an electron pair in the bond formed by atomic orbitals K and L, this operator is in turn an antisymmetrized product of two particle annihilation operators.

$$\Delta_{KL}^{\alpha\beta} = a_{K\alpha} a_{L\beta} - a_{K\beta} a_{L\alpha} \tag{7}$$

At this point it is useful to introduce some properties of $\Delta_{KL}^{\alpha\beta+}$ and $\Delta_{KL}^{\alpha\beta}$, which can be deduced straightforwardly from the well-known anticonmutation properties of the particle creation and annihilation operators.[27]

$$(i) \quad \Delta_{KL}^{\alpha\beta+}\Delta_{MN}^{\alpha\beta+} = \Delta_{MN}^{\alpha\beta+}\Delta_{KL}^{\alpha\beta+}, \quad \Delta_{KL}^{\alpha\beta}\Delta_{MN}^{\alpha\beta} = \Delta_{MN}^{\alpha\beta}\Delta_{KL}^{\alpha\beta}$$

$$(ii) \quad \Delta_{KL}^{\alpha\beta}\Delta_{KM}^{\alpha\beta+}|0\rangle = 0, \quad \Delta_{KL}^{\alpha\beta}\Delta_{MN}^{\alpha\beta+}|0\rangle = 0$$

From property (i) it can be stated that, contrary to the particle creation/annihilation operators, interchange of two $\Delta_{KL}^{\alpha\beta+}$ or $\Delta_{KL}^{\alpha\beta}$ operators does not have any effect on the sign of $|\hat{K}_i\rangle$ or $\langle \hat{K}_i|$, respectively. Therefore, the energy of the system can be written as in eq 8 independently of the order in which the operators $\Delta_{KL(j,n)}^{\alpha\beta+}$ or $\Delta_{KL(i,n)}^{\alpha\beta}$ have been introduced for the construction of the Kekulé operators $\hat{K}_j^+$ or $\hat{K}_i$.

$$E = \frac{1}{n_K} \sum_{i=1}^{n_K} \sum_{j=1}^{n_K} \langle 0 | \hat{K}_i \hat{H} \hat{K}_j^+ | 0 \rangle \tag{8}$$

The transformation of a Kekulé structure into one of its conjugated circuits and vice versa can be also represented using $\Delta_{KL}^{\alpha\beta+}$ and $\Delta_{KL}^{\alpha\beta}$ operators. Thus, we define the operator $\hat{C}_L$ associated to the circuit L as the product of the $\Delta_{KL}^{\alpha\beta}$ operators required to reduce the ket $|K_i\rangle$ to the conjugated circuit L, which can be represented by the ket $|L\rangle$ or $|L'\rangle$ for its counterpart, and $\hat{C}_L^+$ as its adjoint. Obviously, the number of $\Delta_{KL}^{\alpha\beta}$ operators forming $\hat{C}_L$ depends on the conjugated circuit size and the number of π electrons, N_e. Moreover, a conjugated circuit L can have several operators $\hat{C}_L$ associated. This fact is illustrated in Figure 2 for one of the conjugated circuits of phenantrene.

The energy contribution from the superposition of a conjugated circuit L and its counterpart L' can then be obtained from eq 8 by introducing properly the operators associated to L. Taking into account that the product of a Kekulé structure by itself represents only contributions from localized bonds, the energy contribution due to electron conjugation in the circuit L comes exclusively from the crossed products of K_i and K_j structures and then can be written as in eq 9:

$$E_L = \frac{2}{n_K} \sum_{p=1}^{N_L} \sum_{i=1}^{n_K-1} \sum_{j>i}^{n_K} \langle 0 | \hat{K}_i \hat{C}_{L_p}^+ \hat{H} \hat{C}_{L_p} \hat{K}_j^+ | 0 \rangle \tag{9}$$

where the first summation runs over all the operators associated to the conjugated circuit L, the number of operators associated to L is denoted by N_L.

Application of property (ii) reduces drastically the number of non-null terms in eq 9. Thus, if the $\hat{C}_{L_p}$ operator tries to remove a bond electron pair, which does not exist in $\hat{K}_j^+$, the result of applying $\hat{C}_{L_p}$ is zero, the same applies to $\hat{C}_{L_p}^+$ and $\hat{K}_i$ operators. Then, two conditions must be satisfied for a term to be non-null: (1) the operator $\hat{C}_{L_p}$ must contain only bond electron pairs that appear in $\hat{K}_j^+$ and (2) $\hat{K}_i$ and $\hat{K}_j^+$ must differ only in the arranging of the bond electron pairs within the conjugated circuit L. It is straightforward to show that counting the number of terms satisfying conditions (1) and (2) is equal to counting the number of L conjugated circuits, n_L (after taking into account the multiplicative factor 2 in eq 9). Equation 9 then transforms into eq 10:

$$E_L = \frac{1}{2}\frac{n_L}{n_K}[\langle L'|\hat{H}|L\rangle + \langle L|\hat{H}|L'\rangle] = \frac{n_L}{n_K}\langle L'|\hat{H}|L\rangle \tag{10}$$

which represents the stabilization energy due, exclusively, to the electron conjugation within the circuit L. In order to simplify the expressions, we will employ an abbreviated notation for the integrals $\langle L'|\hat{H}|L\rangle$, that will be denoted by $H_{LL'}$ from now on. The total energy due to cyclic conjugation is then the summation of the contributions of all the conjugated circuits (eq 11):

$$E_{CC} = \sum_{L=1}^{n_C} \frac{n_L}{n_K} H_{LL'} = \sum_{L=1}^{n_C} c_L H_{LL'} \tag{11}$$
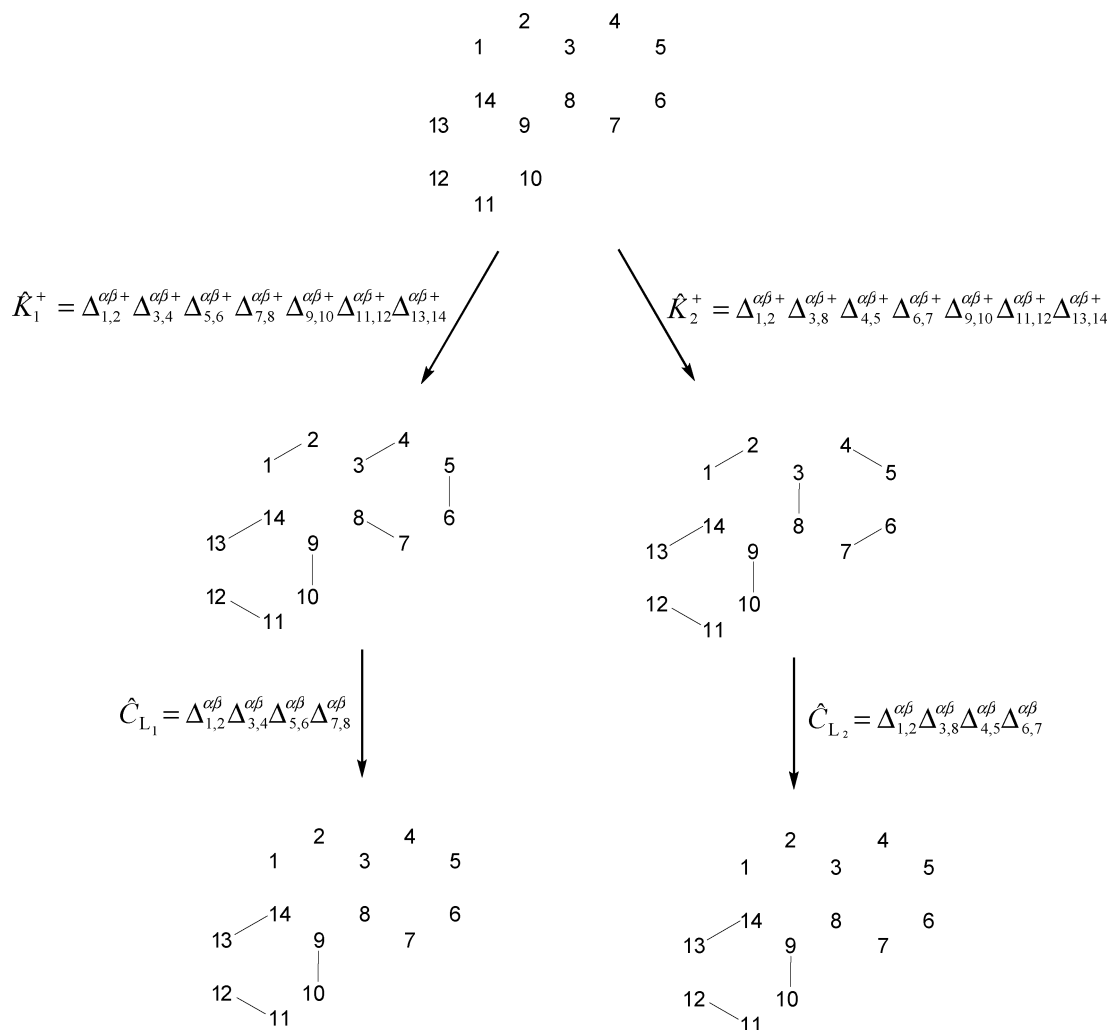
Figure 2. Schematic representation of the effect of some Kekulé operators and conjugated circuit operators of phenantrene.

where n_c is the number of different conjugated structures that results from the superposition of the Kekulé structures. It must be noticed that eq 11 is in fact the conjugated circuit resonance energy in the Randić's model.[1]

Diamagnetic Susceptibility and Ring Current Intensity. It is well-known that the application of a perpendicular magnetic field, H, changes the conjugated circuit resonance energy in a PAH. Obviously, this change affects exclusively the integrals $H_{LL'}$ in eq 11 as the coefficients c_L depend on the number of L conjugated circuits and Kekulé structures, which are not modified by the field. Then, the question that arises is how to account for the effect of a magnetic field over the energy terms $H_{LL'}$ when the field is applied perpendicularly to the molecular plane.

Aihara, in its graph theoretical interpretation of the London theory of diamagnetism, showed that the field-dependent cyclic conjugation energy can be represented by the product of the unperturbed energy with a cosine function of the magnetic field intensity and the area enclosed by the cycle, S_L (see eq 12).[17] Aihara derived this result using the Hückel molecular orbital theory. However, a correct perturbation treatment of the system should provide similar results using both the molecular orbital and valence bond theories. Therefore, we will use the Aihara's result within our valence

bond approach, expressing the field-dependent cyclic conjugation energy by eq 12:

$$E_{CC}(H) = \sum_{L=1}^{n_C} c_L H_{LL'} \cos\left(\frac{2\pi e}{hc} S_L H\right) \quad (12)$$

In the limit of zero magnetic field, the cosine function can be expanded as a Taylor series of H:

$$E_{CC}(H) = \sum_{L=1}^{n_C} c_L H_{LL'}\left(1 - \frac{4\pi^2 e^2}{h^2 c^2} S_L^2 H^2\right) \quad (13)$$

Taking into account that the diamagnetic susceptibility is the second derivative of the energy with respect to the magnetic field, one arrives to eq 14 for the London susceptibility:

$$\chi_d = -\frac{4\pi^2 e^2}{h^2 c^2} \sum_{L=1}^{n_C} c_L H_{LL'} S_L^2 \quad (14)$$

Using the London susceptibility of benzene as reference, eq 14 turns into eq 15:

$$\chi_d = \chi_{d,0} \sum_{L=1}^{n_C} c_L \frac{H_{LL'}}{H_0} \frac{S_L^2}{S_0^2} \qquad (15)$$

where the reference values are denoted by the subindex 0. The contribution of a circuit L is then represented by eq 16:

$$\chi_{d,L} = \chi_{d,0} c_L \frac{H_{LL'}}{H_0} \frac{S_L^2}{S_0^2} \qquad (16)$$

A further simplification can be done for polybenzenoids, if one considers that the area enclosed by the benzene rings is approximately the same as that of the isolated benzene molecule. Thus, the quotient of ring areas can be replaced by the number of benzene rings enclosed by the circuit L, f_L.

$$\chi_{d,L} = \chi_{d,0} c_L \frac{H_{LL'}}{H_0} f_L^2 \qquad (17)$$

The induced electron current intensity per magnetic field unit at the circuit L (eq 18) can be obtained using the well-known relation between the current intensity and the diamagnetic susceptibility (eq 19):[28]

$$I_L = I_0 c_L \frac{H_{LL'}}{H_0} f_L \qquad (18)$$

$$\chi_d H = IS \qquad (19)$$

3. Results and Discussion

The problem with the application of eq 11 is the determination of the terms $H_{LL'}$. By definition, H_0 is the resonance energy of benzene, and then it can be straightforwardly obtained using, for instance, the method of Dewar.[29] However, the calculation of $H_{LL'}$ in polycyclic circuits is not so clear. Within the CCM of Randić, the summation in eq 11 is truncated to circuits containing up to three benzene rings, and then the parameters, called R_1, R_2 and R_3, are calculated via multivariate linear regression using a set of previously computed resonance energies. Values of R_1, R_2 and R_3 can be looked up in reference[1] and compared among different linear regressions obtained using one (R_1), two (R_1 and R_2), and three (R_1, R_2 and R_3) parameters. Comparison of the R_1, R_2 and R_3 values allows concluding that the resonance integrals $H_{LL'}$ decrease proportionally to the circuit size, suggesting that eq 11 may be reformulated in terms of the circuit size. Such reformulation of eq 11 can be done using either the number of rings or the number of centers. We have performed an initial test by using eq 20 as an approximation and also replacing the number of rings, f_L, by the number of centers composing the circuit. In both cases, the results have given rise to similar numerical correlations, although for the calculation of magnetic susceptibilities f_L has provided slightly better results. Thus, we will discuss only the results obtained using the number of rings as a measure of the circuit size:

$$E_{CC} \approx \sum_{L=1}^{n_C} c_L \frac{H_0}{f_L^b} \qquad (20)$$
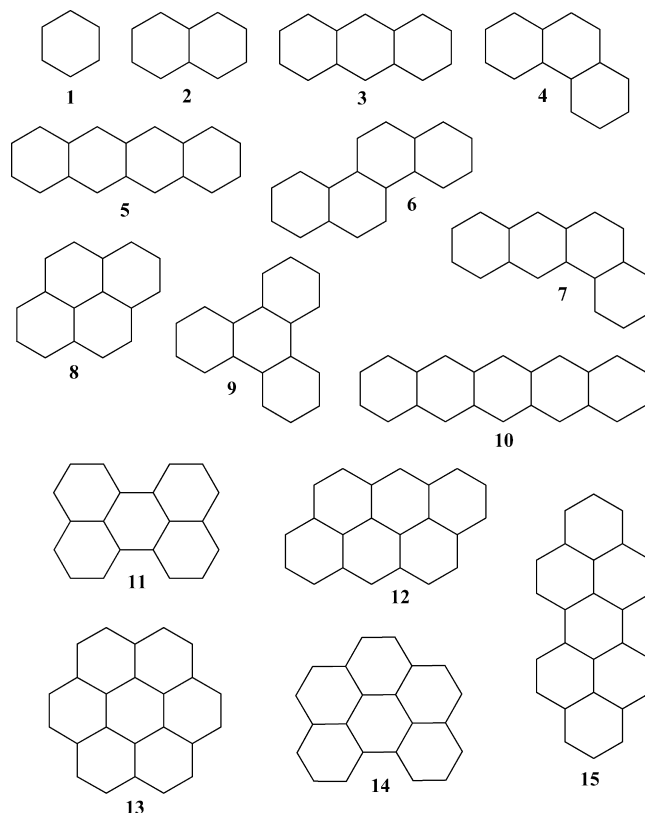
Figure 3. Set of PAHs studied.

Table 1. Dewar and Conjugated Circuits Resonance Energies

molecule	E_{CC}	$E_{CC}{}^a$	DRE[b]	diff[c]
1	0.869	0.869	0.869	0.000
2	1.304	1.304	1.323	−0.020
3	1.569	1.600	1.600	0.000
4	1.950	1.950	1.933	0.017
5	1.750	1.814	1.822	−0.008
6	2.506	2.506	2.483	0.023
7	2.305	2.323	2.291	0.032
8	2.110	2.110	2.098	0.012
9	2.732	2.732	2.654	0.078
10	1.882	1.975	2.004	−0.029
11	2.607	2.607	2.619	−0.012
12	2.621	2.646	2.650	−0.004
13	3.591	3.591	3.524	0.067
14	3.146	3.146	3.128	0.003
15	3.402	3.416	–	–

[a] Including parameter a. [b] Values taken from ref 29. [c] Difference between values listed in the third and fourth columns.

In eq 20, the energy terms $H_{LL'}$ are replaced by the quotient of H_0 and the number of benzene rings of the circuit L, f_L, powered to a parameter, b, whose optimal value can be obtained by statistical regression. Using eq 20, the number of parameters employed reduces to just one since H_0 can be taken as the resonance energy of benzene, truncation to one-two- and three-ring terms is not necessary for the statistical treatment. We have employed eq 20 to reproduce the Dewar resonance energies, DRE, for the series of fifteen PAHs depicted in figure 3. The estimated values are highly close to the correct values when $b = 2$ (see Table 1), the very good linear correlation is reflected on the following statistical parameters: $r = 0.998$, $s = 0.046$, and $F = 2\,987$. However, a look in more detail to the results revels that resonance

Conjugated Circuits

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2699**

Figure 4. E_{CC} vs DRE for the series of PAHs depicted in Figure 2.

Table 2. Magnetic Susceptibilities, Relative to Benzene, For the Set of PAHs Drawn in Figure 1 Estimated Using Conjugated Circuits, Calculated with the London–Hückel Theory and Calculated with The GIAO Method at the B3LYP/6-31G(d,p) Level

molecule	χ_{CC}	$\chi_{CC}{}^a$	$\chi_{Lon}{}^b$	$\chi_\perp$
1	1.00	1.00	1.00	1.00
2	2.00	2.00	2.22	1.83
3	3.00	3.32	3.53	2.67
4	3.20	3.20	3.31	2.55
5	4.00	4.77	4.88	3.53
6	4.50	4.50	4.61	3.34
7	4.29	4.47	–	3.33
8	4.44	4.44	4.70	3.26
9	4.33	4.33	4.33	3.13
10	5.00	6.29	6.26	4.40
11	4.00	4.00	4.17	3.10
12	6.20	6.46	–	4.41
13	10.20	10.20	10.13	5.94
14	6.71	6.71	–	4.69
15	9.00	9.14	–	5.52

^{a} Including parameter *a*. ^{b} Values taken from ref 32.

energies of linear polyacenes are slightly underestimated using eq 20, being more significant with the size of the molecule. This indicates that linear polyacenes display specific behavior, which becomes more noticeable in the calculation of magnetic susceptibilities due to the large circuit size dependence of this magnitude.

The main problem of eq 20 is that it does not distinguish between linear and nonlinear circuits. For instance, eq 20 considers that anthracene and phenantrene, like circuits, have the same $H_{LL'}$ value, but it is known that the contributions to the resonance energy of these circuits is slightly different,[9] being larger in the case of the anthracene circuits. Equation 20 is not able to compensate for these differences in the case of linear polyacenes, where all circuits are linear circuits. Therefore, we have introduced a new parameter in eq 20, which is allowed to adopt different values for nonlinear circuits and linear circuits containing three or more benzene rings.

$$E_{CC} \approx \sum_{L=1}^{n_C} c_L \frac{H_0}{(af_L)^b} \qquad (21)$$

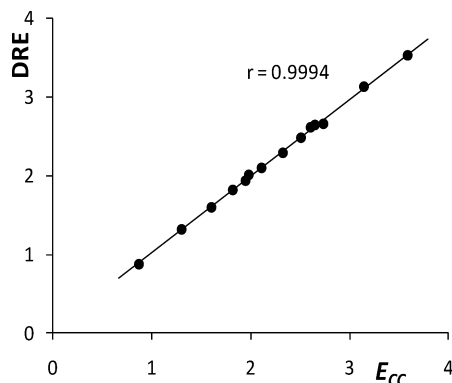This new parameter, *a*, provides more flexibility to the initial expression. Our numerical test has proven that more sophisticated expressions including more parameters are not necessary for obtaining accurate results, both in the case of resonance energies and magnetic susceptibilities. Thus, the best linear fitting between the calculated and estimated values of DRE has been obtained for $b = 2$ and $a = 1$ (nonlinear circuits) and $a = 0.78$ (linear circuits). The perfect correlation obtained is shown in figure 4 with the following statistical parameters: $r = 0.9994$, $s = 0.025$ and $F = 10\,083$. Moreover, as can be seen in Table 1, the largest difference between calculated and estimated values is 0.078, which corresponds to an error lower than 3%.

In light of these results, it can be glimpsed that behind the eqs 20 and 21 there must be a theoretical foundation besides a good statistical correlation. According to them, the aromatic stabilization of a polybenzenoid is a function of the aromatic stabilization of benzene, an idea that was first coined and qualitatively verified by Clar.[30,31] The connection between the aromatic stabilization of both benzene and polybenzenoid is clear in eq 20, the latter depending on the molecular size and the number and shape of the Kekulé structures.

From eq 21 and substituting the optimal value of *b*, $\chi_{d,L}$, and I_L can also be expressed in terms of the coefficient c_L and the number of benzene rings f_L, using benzene as reference.

$$\chi_{d,L} \approx c_L \frac{\chi_{d,0}}{a^2} \qquad (22)$$

$$I_L \approx \frac{c_L}{f_L} \frac{I_0}{a^2} \qquad (23)$$

It is interesting to go over the shape of eq 22. According to it, the contribution of a circuit *L* to the London susceptibility is just given by the quotient between the number of *L* conjugated circuits and the number of Kekulé structures, only in the case of linear circuits it must be corrected with the factor a^2. Therefore, the London susceptibility can be estimated for any polybenzenoid by counting its Kekulé structures and conjugated circuits. The ring codes in the Randić's model[1] appear to be a measure of the circuit contributions to the London susceptibility.

The total London susceptibility results from the summation of all the circuit contributions. In table 2 the values of the magnetic susceptibility calculated using conjugated circuits (with and without the correction parameter *a*), the Hückel–London theory[32] and the GIAO method[33] (at the B3LYP/6-31G(d) level) are collected for the series of PAHs depicted in Figure 3. For the latter, only the component perpendicular to the molecular plane is shown. The magnetic susceptibilities obtained from conjugated circuits are very similar to those calculated with the Hückel–London theory. Only if the correction parameter *a* is not included, the linear polyacenes display specific behavior with significant deviations in the case of molecules 5 and 10. This is completely solved with the inclusion of *a*, as one can see in Figure 5a, the London diamagnetic susceptibilities and the diamagnetic susceptibilities calculated from conjugated circuits display a perfect linear correlation with the following statistical parameters: $r = 0.9991$, $s = 0.115$, and $F = 4\,034$. On the other hand,
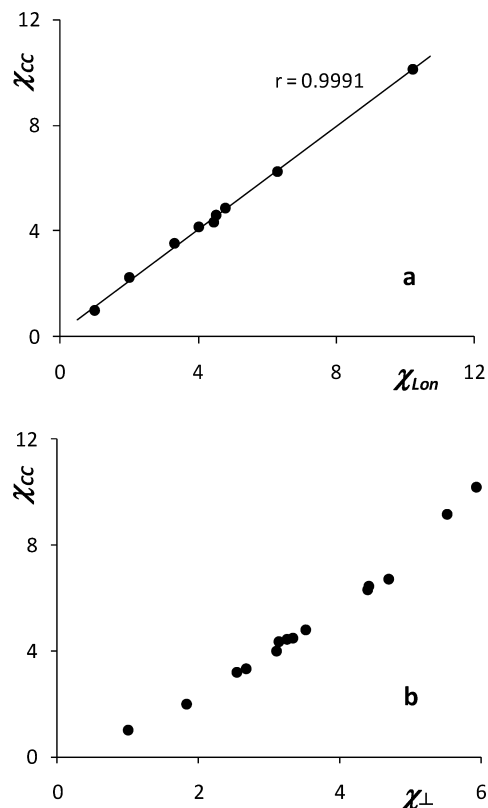
Figure 5. Plots of the magnetic susceptibility estimated using conjugated circuits versus the London diamagnetic susceptibility (a) and the magnetic susceptibility calculated using the GIAO method (b) (only the component perpendicular to the molecular plane).

the comparison with the magnetic susceptibilities calculated with the GIAO method reflects also a good correlation, although in this case, the best fitting equation corresponds to a second-order polynomial instead of a straight line (see Figure 5b). Although magnetic susceptibilities calculated at ab initio level contain also local and paramagnetic contributions, its correlation with the values obtained from conjugated circuits is quite good.

The total induced current intensity I circulating by a given PAH results from the superposition of the individual I_L values. By superimposing the corresponding I_L values calculated using eq 23, we have reconstructed the ring current maps of 14 PAHs (see Figure 6). This set of PAHs comprises a wide range of examples of quite different behavior, which is necessary in order to get a rigorous test of the model and their eventual deficiencies. Moreover, the reader can compare the ring current maps obtained from conjugated circuits with those computed using the ipsocentric method at ab initio level since the latter have been calculated by Fias et al. in ref 34. As one can see, the pictures obtained from conjugated circuits are remarkably accurate. Not only the relative intensity circulating by different bonds in the individual molecules coincides with those computed with the ipsocentric method, but also the relative intensities among different molecules are in very good agreement. Numerical values for the coefficients c_L as well as the current intensities circulating on each
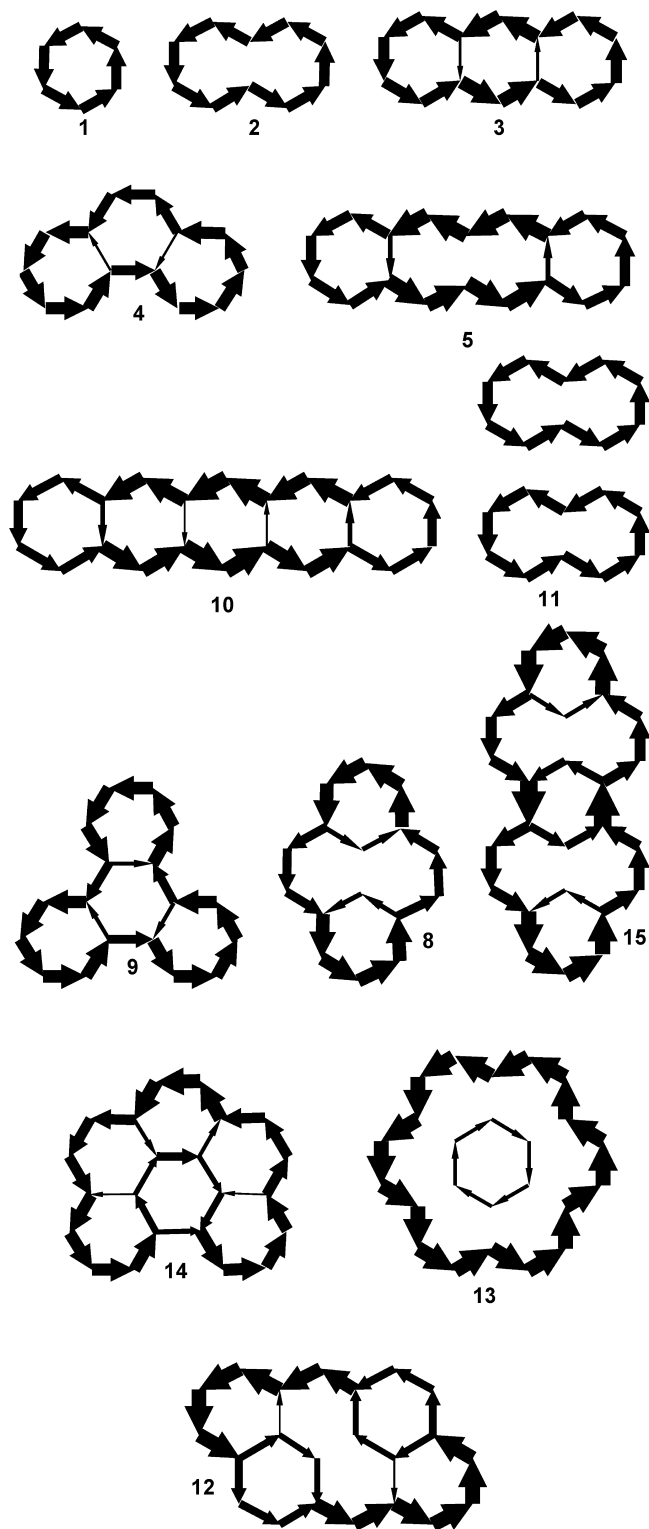
Figure 6. Ring current maps obtained using conjugated circuits.

bond (relative to their values in benzene) can be looked up in the Supporting Information.

We would like to end this discussion pointing out the following statement, which is also an important application. Given that there exist algorithms available for both the counting of the Kekulé structures and conjugated circuits,[11] the calculation of the London susceptibility and the construction of the ring current maps can be under-

Conjugated Circuits

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2701**

taken even for giant PAHs where the ab initio calculations are not feasible. Thus, it is possible to get a quantitative picture of the magnetic properties in giant PAHs with no computational cost.

4. Concluding Remarks and Future Prospects

The conjugated circuits resonance energies have been derived using an operator formalism developed in the context of second quantization. Then, expressions for the London susceptibility and the ring current intensity have been obtained. These expressions have been employed to accurately reproduce the aromatic stabilization, London susceptibilities, and ring current maps of a series of polybenzenoids.

The relevance of the conjugated circuit concept is reflected on the accuracy with that quantum chemical results are reproduced, providing quantitative information not only about the aromatic stabilization of a PAH but also about its magnetic behavior. This methodology can be applied on both small and medium size PAHs in combination with more accurate ab initio methods, providing a local interpretation of the results, but it can also be applied on giant PAHs to get accurate pictures of their magnetic behaviors when ab initio calculations are not applicable.

Calculations only require the previous knowledge of the resonance energy of benzene by any of the methods available, and then the remaining properties are obtained for larger polybenzenoids with no computational cost. The next step is to generalize the method for other PAHs containing five-center and four-center conjugated rings by introducing some additional parameters.

Supporting Information Available: Graphs representing the conjugated circuits, c_L values, and current intensities circulating through the bonds relative to benzene. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Randić, M. *Chem. Rev.* **2003**, *103*, 3449.

(2) Randić, M. *Chem. Phys. Lett.* **1976**, *38*, 3839.

(3) Randić, M. *Tetrahedron* **1977**, *33*, 1905.

(4) Randić, M. *J. Am. Chem. Soc.* **1977**, *99*, 444.

(5) Gomes, J. A. N. F.; Mallion, R. B. *Rev. Port. Quim.* **1979**, *21*, 82.

(6) Simpson, W. T. *J. Am. Chem. Soc.* **1953**, *75*, 597.

(7) Herndon, W. C. *J. Am. Chem. Soc.* **1973**, *95*, 2404.

(8) Herndon, W. C.; Ellzey, M. L., Jr. *J. Am. Chem. Soc.* **1974**, *96*, 6631.

(9) Klein, D. J.; Trinajstić, N. *Pure Appl. Chem.* **1989**, *61*, 2107.

(10) Pauling, L.; Wheland, G. W. *J. Chem. Phys.* **1933**, *1*, 362.

(11) See ref 1 and references therein.

(12) Randić, M. *Pure Appl. Chem.* **1980**, *52*, 1587.

(13) London, F. *J. Phys. Radium* **1937**, *8*, 397.

(14) Aihara, J.-I. *J. Am. Chem. Soc.* **1979**, *101*, 558.

(15) Aihara, J.-I. *J. Am. Chem. Soc.* **1979**, *101*, 5913.

(16) Aihara, J.-I. *J. Am. Chem. Soc.* **1981**, *103*, 5704.

(17) Aihara, J.-I. *J. Am. Chem. Soc.* **1979**, *101*, 558.

(18) Aihara, J.-I.; Horikawa, T. *Bull. Chem. Soc. Jpn.* **1983**, *56*, 1853.

(19) Aihara, J.-I. *J. Am. Chem. Soc.* **1976**, *98*, 2750.

(20) Gutman, I.; Milun, M.; Trinajstić, N. *J. Am. Chem. Soc.* **1977**, *99*, 1692.

(21) Aihara, J.-I. *Pure Appl. Chem.* **1982**, *54*, 1115.

(22) Aihara, J.-I. *J. Am. Chem. Soc.* **2006**, *128*, 2873.

(23) Aihara, J.-I. *Bull. Chem. Soc. Jpn.* **2004**, *77*, 651.

(24) Hückel, E. *Z. Physik* **1931**, *71*, 204.

(25) Hückel, E. *Z. Physik* **1932**, *76*, 628.

(26) Cantu, A. A.; Klein, D. J.; Matsen, F. A. *Pure Appl. Chem.* **1975**, *38*, 341.

(27) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; McGraw-Hill, New York, 1989.

(28) Pople, J. A.; Untch, K. G. *J. Am. Chem. Soc.* **1966**, *88*, 4811.

(29) Dewar, M. J. S.; De Llano, C. *J. Am. Chem. Soc.* **1969**, *91*, 789.

(30) Clar, E. *The Aromatic Sextet*; J. Wiley & Sons: London, 1972.

(31) Clar, E. *Polycyclic Hydrocarbons*, Academic Press and Springer: London and Berlin, Vols. 1 and 2; 1964.

(32) Pullman, B.; Pullman, A. *Les Theories Electronique de la Chimie Organique*, Masson et Cie.: Paris, France, 1952; pp 545.

(33) Wolinski, K.; Hilton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.

(34) Fias, S.; Fowler, P. W.; Delgado, J. L.; Hahn, U.; Bultinck, P. *Chem.—Eur. J.* **2008**, *14*, 3093.

Block-Localized Density Functional Theory (BLDFT), Diabatic Coupling, and Their Use in Valence Bond Theory for Representing Reactive Potential Energy Surfaces

Alessandro Cembran,[†] Lingchun Song,*,[†] Yirong Mo,*, [‡] and Jiali Gao*,[†]

Department of Chemistry, Digital Technology Center and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455, and Department of Chemistry, Western Michigan University, Kalamazoo, Michigan 49008

Abstract: A multistate density functional theory in the framework of the valence bond model is described. The method is based on a block-localized density functional theory (BLDFT) for the construction of valence-bond-like diabatic electronic states and is suitable for the study of electron transfer reactions and for the representation of reactive potential energy surfaces. The method is equivalent to a valence bond theory with the treatment of the localized configurations by using density functional theory (VBDFT). In VBDFT, the electron densities and energies of the valence bond states are determined by BLDFT. A functional estimate of the off-diagonal matrix elements of the VB Hamiltonian is proposed, making use of the overlap integral between Kohn−Sham determinants and the exchange-correlation functional for the ground state substituted with the transition (exchange) density. In addition, we describe an approximate approach, in which the off-diagonal matrix element is computed by wave function theory using block-localized Kohn−Sham orbitals. The key feature is that the electron density of the adiabatic ground state is not directly computed nor used to obtain the ground-state energy; the energy is determined by diagonalization of the multistate valence bond Hamiltonian. This represents a departure from the standard single-determinant Kohn−Sham density functional theory. The multistate VBDFT method is illustrated by the bond dissociation of H_2^+ and a set of three nucleophilic substitution reactions in the DBH24 database. In the dissociation of H_2^+, the VBDFT method yields the correct asymptotic behavior as the two protons stretch to infinity, whereas approximate functionals fail badly. For the S_N2 nucleophilic substitution reactions, the hybrid functional B3LYP severely underestimates the barrier heights, while the approximate two-state VBDFT method overcomes the self-interaction error, and overestimates the barrier heights. Inclusion of the ionic state in a three-state model, VBDFT(3), significantly improves the computed barrier heights, which are found to be in accord with accurate results. The BLDFT method is a versatile theory that can be used to analyze conventional DFT results to gain insight into chemical bonding properties, and it is illustrated by examining the intricate energy contributions to the ion−dipole complex stabilization.

1. Introduction

The construction of an effective two-state diabatic Hamiltonian provides an extremely useful tool for many chemical

and biochemical applications.[1−5] Examples include the study of electron transfer (ET) using electronic structure methods and the representation of the potential energy surface (PES) of chemical reactions in condensed phases and in enzymes. Although these applications appear to be very different, the underlying principles to determine the PES are identical. In the Marcus−Hush electron transfer theory, the coupling

* Corresponding author e-mail: songx184@umn.edu (L.S.); yirong.mo@wmich.edu (Y.M.); gao@jialigao.org (J.G.).

† University of Minnesota.

‡ Western Michigan University.

matrix element between donor (H_D) and acceptor (H_A) potential energy surfaces, V_{DA}, is a crucial element in the calculation of the electron-transfer rate.[6] At the crossing point where $H_D = H_A$, the coupling strength is the difference between the diabatic state energy and the adiabatic ground-state energy (ε_1):

$$V_{DA} = H_A - \varepsilon_1 = \frac{|H_A S_{DA} - H_{DA}|}{1 + S_{DA}} \qquad (1)$$

where $H_{DA} = <\Phi_D|H|\Phi_A>$ and $S_{DA} = <\Phi_D|\Phi_A>$ are, respectively, the exchange and overlap integrals of the donor and acceptor valence-bond electronic wave functions, Φ_D and Φ_A, and H is the electronic Hamiltonian of the system.

For chemical reactions, if Φ_1 and Φ_2 represent the valence bond electronic wave functions of the reactant and product states,[7] the Born–Oppenheimer PES of the adiabatic ground state is given by the lower root, ε_g, of the 2×2 generalized secular equation:

$$\begin{vmatrix} H_{11} - \varepsilon & H_{12} - \varepsilon S_{12} \\ H_{21} - \varepsilon S_{21} & H_{22} - \varepsilon \end{vmatrix} = 0 \qquad (2)$$

where the Hamiltonian matrix element is defined by $H_{ij} = <\Phi_i|H|\Phi_j>$, S_{ij} is the overlap integral, and H is the total effective Hamiltonian that includes both solute and solvent contributions. In this discussion, the electronic coupling may be defined as[4]

$$V_{12} = H_{12} - \varepsilon_g S_{12} \qquad (3)$$

Then, the adiabatic ground-state potential energy has the simple form

$$\varepsilon_g = \frac{1}{2}(H_{11} + H_{22}) - \frac{1}{2}[(H_{11} - H_{22})^2 + 4V_{12}^2]^{1/2} \qquad (4)$$

The electronic coupling matrix element V_{12} may also be back-calculated from the diabatic and adiabatic ground-state potential energies:

$$V_{12} = \sqrt{(H_{11} - \varepsilon_g)(H_{22} - \varepsilon_g)} \qquad (5)$$

The two expressions of V_{12} in eqs 3 and 5 are equivalent, shown to emphasize an important point. The first shows the explicit dependence of the overlap integral, whereas this dependence is implicit in eq 5. In either case, there is no need to restrict the diabatic states to be orthogonal.

When the diabatic state energies and the adiabatic ground-state energy are determined at different levels of theory (or from experiments), one assumes that the difference in absolute energy is constant along the reaction path. As such, the adiabatic ground-state energy profile can be shifted by a constant Δ to yield an approximate V_{12}:

$$V_{12} = \sqrt{(H_{11}^{LL} - \varepsilon_g^{HL} + \Delta)(H_{22}^{LL} - \varepsilon_g^{HL} + \Delta)} \qquad (6)$$

where LL and HL indicate energies are computed at lower-level and higher-level theories, and Δ can be chosen with the assumption that there is no diabatic coupling at the reactant state. Importantly, eq 6 provides a convenient connection between the diabatic and adiabatic PESs when

they are obtained from different levels of theory, and hence, V_{12} may be fitted globally or semiglobally to reproduce the all-dimensional adiabatic PES for a given two-state model.[8–12] In this case, the dependence on overlap is implicit. However, the effect of solvation on V_{12} may not be adequately assessed using eq 6 if the diabatic PES (H_{11} and H_{22}) and the adiabatic ground-state energy are not determined consistently with the same basis and theory or by optimizing both the orbital and configuration coefficients simultaneously.[3,13–16]

The goal of this work is to develop a multistate density functional theory (MSDFT) in the framework of the valence bond model, in which the matrix elements are determined by Kohn–Sham (KS) density functional theory (DFT). In this way, dynamic electron correlation effects are treated by DFT for the diabatic (VB) states, whereas static electron correlation is partially described by the multiconfigurational valence bond Hamiltonian (keeping in mind the possibility of double-counting correlation effects using the current functionals). Since diabatic states are not uniquely defined,[3,4,7,17–20] the problem in these applications involves both the construction of the diabatic states and the accurate computation of the matrix elements.

Numerous methods have been developed.[2,5,6,21–30] To this end, we have developed a mixed molecular orbital and valence bond (MOVB) theory,[3,13,14,16] based on a block-localized molecular orbital method.[31–38] In MOVB, diabatic states are defined as the valence bond configurations of the reactant and product structure, and the electronic wave functions are approximated by a single Slater determinant, in which the molecular orbitals (MOs) are strictly localized within the individual fragments of a molecular configuration. We note that the MOs within each fragment can be orthogonalized, but they are nonorthogonal between different fragments. Consequently, MOVB retains important characteristics of valence bond theory.[39–42] When the system contains just one fragment, MOVB reduces exactly to the conventional Hartree–Fock theory. On the other hand, when MOs are fully localized, MOVB becomes an ab initio valence bond self-consist field model,[41,43,44] which is equivalent to the complete active space (CASSCF) approach.

The single determinant, block-localized molecular orbitals can be used to compute the electron density in exactly the same way as that in the KS DFT approach.[45] Consequently, by construction, the total electron density of a molecular system is block-localized with the desired characteristics, and such a block-localized density functional theory (BLDFT) can be used to obtain the ground-state energy of the corresponding charge and spin-localized VB configuration. The block-localized MOs used in DFT calculations correspond to block-localized Kohn–Sham (BLKS) orbitals, which can be defined analogously as the MOVB definition of the reactant (donor) state and the product (acceptor) state for chemical (ET) reactions.[3,12,13,16] In this article, we describe two methods for computing the off-diagonal matrix elements using density functional theory. The MSDFT method can be adapted in VB calculations, so the method is also called the mixed VB and DFT (VBDFT) method. In this context, MSDFT and VBDFT can be interchangeably used, and we choose to use VBDFT throughout the rest of

the paper. Clearly, an advantage of using BLDFT in VB calculations is the inclusion of electron correlation effects. Furthermore, because VBDFT is a multiconfigurational method, the self-interaction error in the existing functionals can, in principle, be eliminated in bond-making and -breaking processes.

For comparison, a related but distinct method is the constrained DFT (CDFT) formulated by Dederichs et al. in 1984,[46] in which a set of charge and spin constraints is introduced into KS-DFT calculations by the method of a Lagrange multiplier. The CDFT method has been used in a number of applications. Scheffler and co-workers studied the adsorption of triplet oxygen on the aluminum surface and reactions that are forbidden by the spin selection rule.[47,48] Van Voorhis described an approach to estimate the coupling matrix element in long-range ET reactions and used it in configuration interaction (CI) with CASSCF-inspired charge- and spin-constrained configurations.[29,30,49] Wesolowski developed a constrained DFT algorithm to describe intermolecular interactions.[50,51] In contrast, the VBDFT method follows a *different strategy*. We block-localize the KS orbitals, and hence the associated electron density is constrained by construction (e.g., the Mulliken population constraint is inherently imposed). The electron density for each diabatic state is derived from an antisymmetric wave function, consisting of both orthogonal and nonorthogonal BLKS orbitals, a feature distinct from the CDFT approach of Dederichs et al.[46] but characteristic of VB theory.[7] Notably, the VBDFT method is applicable to both short-range (or strong-coupling) and long-range (or weak-coupling) ET reactions.

In the following, we first present the theoretical background of the VBDFT (or MSDFT) method and computational details. Then, we illustrate the VBDFT method by considering a set of applications involving S_N2 reactions. The paper is concluded with a summary of major findings and future perspectives.

2. Methods

Often, it is desirable to impose certain charge and spin constraints within a molecular system to describe the localized chemical bonding character to gain insight into properties such as resonance and charge transfer, concepts central to the understanding of the chemical bond. The MOVB method was developed to provide a computationally efficient procedure to study these effects as well as chemical reactions in solution and in enzymes.[3,13−16,52] In this section, we present an approach to use BLDFT to evaluate the matrix elements of VB Hamiltonians. Thus, this multireference density functional approach is also called the VBDFT method, which represents an extension of the MOVB model to include dynamic electron correlation effects in ab initio VB-like calculations.

Throughout this paper, we use the following convention. Subgroups of the molecular system are specified by Roman capital letters A, B, ...; (block-localized) Kohn−Sham molecular orbitals are represented by $|\psi\rangle$, labeled by the lower case letters, i, j, ...; atomic orbital basis functions are denoted by $|\chi\rangle$, indexed by the lower case Greek letters, μ, ν,... When

A, B, ... are used as superscripts over a matrix (distinguished by bold letters), they denote the dimension of the matrix in terms of the corresponding MOs in that subgroup, whereas when they are used as subscripts, the matrix is defined in terms of the basis functions. The elements of a column (row) in a matrix are always subscripted. VB configurations are specified by u, v, w, ...

A. Block-Localized Density Functional Theory. In BLDFT, we first construct a set of BLKS orbitals corresponding to the desired charge and spin character.[31,32,45] For convenience of discussion, we consider a closed-shell system both for the entire system and for subgroups; generalization to open-shell and spin-constrained cases is straightforward and has been implemented. The full system consists of N electrons and M basis functions that are partitioned into K subgroups with n_A electrons and m_A basis functions in subgroup A. Then, the BLKS orbitals are expressed in terms of the atomic orbitals located on atoms in a particular subspace:

$$|\psi_i^A\rangle = \chi_A \mathbf{C}_i^A = \sum_{\mu=1}^{m_A} |\chi_{A\mu}\rangle c_{\mu i}^A, \quad A = 1, ..., K \quad (7)$$

where $c_{\mu i}^A$ is an element of the column coefficient vector ($\mathbf{C}_i^A$) of KS orbital i ($|\psi_i^A\rangle$), and $\{|\chi_{A\mu}\rangle\}_{\mu=1}^{m_A}$ are the basis functions in subgroup A, arranged as a row vector χ_A. Let $\Omega^A = \psi_1^A \alpha \psi_1^A \beta ... \psi_{n_A/2}^A \beta$ be a successive product of n_A occupied spin−orbitals in subgroup A, and α and β are spin functions. The Slater determinant function for the block-localized system is constructed as follows:

$$\Phi_u = N_u \hat{A}\{\Omega^1 \Omega^2 ... \Omega^K\} \quad (8)$$

where $\hat{A}$ is an antisymmetrization operator, and N_u is the normalization constant.

The BLKS orbitals between different subgroups in eq 8 are not orthogonal. Orthogonalization of these orbitals can be done and is in fact part of the procedure to solve the self-consistent field equations; however, the orthorgonalized orbitals should not be used to interpret the properties of the strictly localized, individual blocks or separate molecules because they contain orthogonalization tails. Note that the KS orbitals within the same subgroup can be and are orthorgonalized, which does not affect the total energy by a unitary transformation. The overlap matrix of the MOs is given as follows

$$\mathbf{S} = \mathbf{C}^T \mathbf{R} \mathbf{C} \quad (9)$$

where $\mathbf{R} = \chi^T \chi$ is the overlap in terms of the basis functions and $\mathbf{C}$ is the transformation matrix:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}^1 & 0 & \cdots & 0 \\ 0 & \mathbf{C}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{C}^K \end{pmatrix} \quad (10)$$

The one-particle density matrix from the occupied nonorthogonal BLKS orbitals is defined by

BLDFT and Diabatic Coupling

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2705**

$$\mathbf{D} = \mathbf{C}(\mathbf{C}^T\mathbf{R}\mathbf{C})^{-1}\mathbf{C}^T \tag{11}$$

which satisfies the symmetry ($\mathbf{D}^T = \mathbf{D}$), rank ($Tr(\mathbf{DR}) = N$), and idempotency ($\mathbf{DRD} = \mathbf{D}$) conditions, and the electron density is given as follows

$$\rho(\mathbf{r}) = \sum_{\mu\nu}^{m} |\chi_\mu(\mathbf{r}) > D_{\mu\nu} < \chi_\nu(\mathbf{r})| = \chi(\mathbf{r})\,\mathbf{D}\chi^T(\mathbf{r}) \tag{12}$$

Using the one-particle density matrix and electron density of eqs 11 and 12 computed from the nonorthogonal KS orbitals, we can express the BLDFT ground-state energy identically to that in the case of orthogonal KS orbitals:[45]

$$E[\rho] = Tr(\mathbf{Dh}) + \frac{1}{2}Tr(\mathbf{DJD}) + E_{xc}[\rho(\mathbf{r})] + E_{nuc} \tag{13}$$

where E_{nuc} is the coulomb energy of the nuclei, $\mathbf{h}$ and $\mathbf{J}$ are the usual Hamiltonian (one-electron) and Coulomb integral matrices, and $E_{xc}[\rho(\mathbf{r})]$ is the exchange-correlation energy functional.

The corresponding block-localized Kohn−Sham equations for the nonorthogonal KS orbitals can be derived following the procedure described by Stoll et al.,[53] and later by a number of groups in various forms in molecular orbital theory,[14,31−38,54] primarily for treating intermolecular interactions without basis set superposition errors and for energy decomposition analysis. Because the transformation matrix (eq 10) is block-diagonal, the conventional self-consistent field (SCF) procedure for the Kohn−Sham equations of the entire system can be cast into K separate KS equations, one set for each subgroup. First, we define the projection operator $\hat{P}_{\notin A}$ in the space of occupied KS orbitals that exclude those in subgroup A, indicated by the subscripts $\notin A$:

$$\hat{P}_{\notin A} = \sum_{B,C\neq A}^{K} \sum_{i,j}^{occ} |\psi_i^B > [(\mathbf{S}_{\notin A})^{-1}]_{ij} < \psi_j^C| \tag{14}$$

where $\mathbf{S}_{\notin A}$ is the overlap matrix without the occupied KS orbitals of subgroup A. Then, the KS equations for the orbitals of subgroup A are given as follows

$$\hat{F}^A|\psi_i^A > \, = (\hat{1} - \hat{P}_{\notin A})|\psi_i^A > \varepsilon_i^A \tag{15}$$

where $\hat{F}^A$ is the projected KS operator

$$\hat{F}^A = (\hat{1} - \hat{P}_{\notin A})\hat{F}(\hat{1} - \hat{P}_{\notin A}) \tag{16}$$

with $\hat{F}$ being the unprojected (conventional) KS operator,

$$\hat{F} = -\frac{1}{2}\nabla^2 + v_{ext}(r) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\,d\mathbf{r}' + v_{xc}[\rho(\mathbf{r})] \tag{17}$$

where $v_{xc}[\rho(\mathbf{r})]$ is the exchange-correlation potential, and $v_{ext}(r)$ is the nuclei−electron attraction, *which may include the partial charges from the solvent in combined QM/MM calculations.*

The KS orbitals in eq 15 can be optimized sequentially by Jacobi rotation,[13,31] which is straightforward to implement. Alternatively, in matrix form, in terms of the basis functions, the generalized secular equations can be written as follows:[14,37,54]

$$\mathbf{F}_{AA}^P\mathbf{C}_{AA} = \mathbf{R}_{AA}^P\mathbf{C}_{AA}\varepsilon_{AA} \tag{18}$$

where ε_{AA} is a diagonal matrix corresponding to orbital energies. Equation 18 was derived subject to the condition (which does not affect the energy):[34]

$$(\mathbf{C}_{AA})^T\mathbf{R}_{AA}^P\mathbf{C}_{AA} = \mathbf{1}_{AA} \tag{19}$$

and the projected overlap ($\mathbf{R}_{AA}^P$) and KS-Fock ($\mathbf{F}_{AA}^P$) matrices are given as

$$\mathbf{R}_{AA}^P = (\mathbf{R}_{AA}, \mathbf{R}_{AB})\mathbf{P}_{\notin A} \tag{20}$$

and

$$\mathbf{F}_{AA}^P = (\mathbf{P}_{\notin A})^T\mathbf{F}\mathbf{P}_{\notin A} \tag{21}$$

The projection matrix is defined below. The computational procedure is conveniently described by considering an effective partition of two blocks, A and B, where the orbitals in A are being optimized and B includes all other subgroups of the molecular system. We arrange block A as the first block and B second; that is, we rearrange the basis in the following order, specified by a prime: $\chi' = [\chi_A, \chi_1, ..., \chi_{A-1}, \chi_{A+1}, ..., \chi_K] = [\chi_A, \chi_B]$, and thus the transformation matrix $\mathbf{C}' = [\mathbf{C}^A, \mathbf{C}^B]$. Then, the overlap, density, and projection matrices in terms of the basis functions and the overlap matrix in terms of KS orbitals in eqs 20 and 21 are given, respectively, as

$$\mathbf{R}' = \chi'^T\chi' = \begin{pmatrix} \mathbf{R}_{AA} & \mathbf{R}_{AB} \\ \mathbf{R}_{BA} & \mathbf{R}_{BB} \end{pmatrix} \tag{22}$$

$$\mathbf{D}_{BB} = \mathbf{C}^B(\mathbf{S}^{BB})^{-1}(\mathbf{C}^B)^T \tag{23}$$

$$\mathbf{P}_{\notin A} = \begin{pmatrix} \mathbf{1}_{AA} \\ -\mathbf{D}_{BB}\mathbf{R}_{BA} \end{pmatrix} \tag{24}$$

$$\mathbf{S}' = \begin{pmatrix} \mathbf{S}^{AA} & \mathbf{S}^{AB} \\ \mathbf{S}^{BA} & \mathbf{S}^{BB} \end{pmatrix} \tag{25}$$

Starting from an initial guess, typically generated from the extended Hückel method but preferably from calculations of individual subgroups, one iteratively optimizes the KS orbitals of each subgroup employing the densities generated previously for other subgroups until the total energy and density are converged. The key property of these orbitals is that they yield the exact electron density of the charge- and spin-constrained configuration through eq 12, which is then used to determine $E_{xc}[\rho(\mathbf{r})]$ in eq 13.[45] Furthermore, as illustrated in Hartree−Fock theory by Nagata et al.,[54] it is straightforward to show that, if the electron density is partitioned in the sense of Mulliken population of the occupied, nonorthogonal BLKS orbitals, the electron densities of the individual subgroups are conserved without charge transfer between different ones. Thus, the total electron density is partitioned into subgroups as follows:

$$N = N^A + N^B = Tr([\mathbf{DR}]_{AA}) + Tr([\mathbf{DR}]_{BB}) \tag{26}$$

where $[\mathbf{DR}]_{AA} = \mathbf{D}_{AA}\mathbf{R}_{AA} + \mathbf{D}_{AB}\mathbf{R}_{BA}$ and $[\mathbf{DR}]_{BB} = \mathbf{D}_{BA}\mathbf{R}_{AB} + \mathbf{D}_{BB}\mathbf{R}_{BB}$. The total electron density is similarly partitioned

2706 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Cembran et al.

into the electron densities of the subgroups

$$\rho(\mathbf{r}) = \rho^A(\mathbf{r}) + \rho^B(\mathbf{r}) \tag{27}$$

which are given by

$$\rho^A(\mathbf{r}) = \chi_A(\mathbf{r}) \, \mathbf{D}_{AA}(\chi_A)^T(\mathbf{r}) + \chi_A(\mathbf{r}) \, \mathbf{D}_{AB}(\chi_B)^T(\mathbf{r}) \tag{28}$$

$$\rho^B(\mathbf{r}) = \chi_B(\mathbf{r}) \, \mathbf{D}_{BB}(\chi_B)^T(\mathbf{r}) + \chi_B(\mathbf{r}) \, \mathbf{D}_{BA}(\chi_A)^T(\mathbf{r}) \tag{29}$$

For subgroup A, the number of partitioned electrons is

$$N^A = \int \rho^A(\mathbf{r}) \, d\mathbf{r} = Tr(\mathbf{D}_{AA}\mathbf{R}_{AA} + \mathbf{D}_{AB}\mathbf{R}_{BA}) = n_A \tag{30}$$

which is the number of constrained electrons in subgroup A. Therefore, the block-localized KS orbital partition defined by eq 8 is equivalent to imposing the constraint, $\delta(A)$, to the total electron density, such that

$$\int \delta(A) \, \rho(\mathbf{r}) \, d\mathbf{r} - n_A = \int \rho^A(\mathbf{r}) \, d\mathbf{r} - n_A = 0 \tag{31}$$

Before we leave this section, it is of interest to make a comparison with constrained DFT (CDFT) introduced by Dederichs et al.,[46] in which the constrained KS equations are

$$[\hat{F} + V_c w_c(\mathbf{r})]|\phi_i^c \rangle = |\phi_i^c \rangle \varepsilon_i^c \tag{32}$$

where V_c is the Lagrangian multiplier for the constraint $w_c(\mathbf{r})$ to yield a desired charge and spin property N_c, $\int w_c(\mathbf{r}) \, \rho(\mathbf{r}) \, d\mathbf{r} - N_c = 0$. In CDFT, the constrained KS orbitals are expanded over the entire basis set, corresponding to a fully delocalized system with specific spatial restrictions (N_c) in practice.[30] In contrast, the present BLDFT builds charge and spin localization by construction, and the charge and spin densities are strictly constrained within each subgroup in the sense of Mulliken population partition without ambiguity or integration uncertainty (however, this should not be confused with the use of the Mulliken population constraint in CDFT calculations, in which case it is difficult to achieve charge localization because of the use of the intrinsically delocalized KS orbitals).[30] Note that the charge integration scheme in CDFT is rather arbitrarily defined and cannot yield the correct total charges for the constrained fragments in the dissociation of H_2^+.

There are, of course, many other ways of representing constrained DFT calculations, including a useful method developed by Wesolowski et al. for studying intermolecular interactions.[51] In this case, the molecular fragments are considered as separated systems under the influence of the mutual polarization of other densities, either frozen or optimized. An anonymous reviewer pointed out that, in principle, the method developed by Wesolowski can be used to construct the localized configurations as described here, provided if one knew the exact nonadditive part of the kinetic energy. "The problem is that the function isn't known and the approximations currently used will fail when there is significant subgroup overlap." Of course, subgroup overlap is significant in problems of interest to us. In this regard, the kinetic energy in the present BLDFT method is treated exactly as that in the KS-DFT scheme, and the resulting

valence bond Hamiltonian corresponds to a multireference density functional theory, in which the densities for the VB-like diabatic (reference) states are uniquely defined.

B. Diabatic Coupling and Multireference Valence Bond Density Functional Theory. In this section, we formulate a two-state VB Hamiltonian, making use of BLDFT to determine the matrix elements. Clearly, the method is general and has been implemented for multiple states as illustrated in the following applications. For the S_N2 reaction between nucleophile Nu^- and substrate CH_3L, we define the reactant (Ψ_u: $[Nu^-][CH_3L]$) and the product (Ψ_w: $[NuCH_3][L^-]$) diabatic states on the basis of the VB structure of separate, but interacting, fragments:[3,13]

$$\Psi_u(\mathbf{X}) = N_u\hat{A}\{\Omega^1[Nu^-] \, \Omega^2[CH_3L]\} \tag{33}$$

$$\Psi_w(\mathbf{X}) = N_w\hat{A}\{\Omega^3[NuCH_3] \, \Omega^4[L^-]\} \tag{34}$$

where N_u and N_w are normalization constants and $\mathbf{X}$ specifies the instantaneous atomic coordinates. Then, the VB wave function is written as a linear combination of the two VB configurations defined by eqs 33 and 34, which is valid for the description of the entire adiabatic ground state and a coupling excited state:

$$\Phi^{VB}(\mathbf{X}) = a_u\Psi_u(\mathbf{X}) + a_w\Psi_w(\mathbf{X}) \tag{35}$$

The adiabatic ground-state energy at a given molecular geometry, $\mathbf{X}$, is the lower root of the generalized secular equation (eq 2), in which $u = 1$ and $w = 2$.

Note that the wave function of eq 35 is used purely for the purpose of deriving the secular equation (eq 2). Importantly, $\Phi^{VB}(\mathbf{X})$ is determined corresponding to a unique potential $v^{VB}(\mathbf{X})$, which in turn defines the adiabatic ground state density and the adiabatic ground-state energy within the block-localized configurations defined by eqs 33 and 34. However, $\Phi^{VB}(\mathbf{X})$ is not used to compute the electron density to determine the DFT energy—a departure from previous approaches employing multiconfigurational wave function to construct the one-particle density. The KS determinants defined in eqs 33 and 34 are used to generate the BLKS orbitals, which yield the exact electron densities for the individual diabatic states and the corresponding matrix elements H_{uu} and H_{ww} (eq 13). Unfortunately, the exact wave functions associated with the block-localized densities for the reactant and product diabatic states are not available. Thus, the exact overlap integral in eq 2 between two diabatic states defined by BLDFT cannot be directly computed. An obvious choice to circumvent this difficulty is to use the KS determinants to evaluate the overlap, S_{uw}. This strategy has been used by Wu and others.[49] Then, the overlap integral between Ψ_u and Ψ_w can be decomposed into the product of overlaps for the α and β spin orbitals.[14] As we are considering closed-shell cases and each BLKS orbital is occupied by two electrons, the overlaps for the α and β spin orbitals are the same and equal to q_{uw}. Thus,

$$S_{uw} = \langle \Psi^u | \Psi^w \rangle = q_{uw}^2 \tag{36}$$

BLDFT and Diabatic Coupling

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2707**

where q_{uw} is the determinant of the overlap matrix between the BLKS orbitals from the two states:

$$\mathbf{S}^{uw} = (\mathbf{C}^u)^{\mathrm{T}} \mathbf{R} \mathbf{C}^w \tag{37}$$

$$q_{uw} = \det |\mathbf{S}^{uw}| \tag{38}$$

where $\mathbf{C}^u$ and $\mathbf{C}^w$ are the transformation matrices of states Ψ_u and Ψ_w which have the same total basis functions and occupied orbitals, but, of course, are block-localized differently.

The coupling matrix element represents the transition between two diabatic states Ψ_u and Ψ_w whose energies are uniquely defined by their electron densities. However, H_{uw} is not a natural functional of the electron density, but it is a functional of both the electron densities $\rho_u(\mathbf{r})$ and $\rho_w(\mathbf{r})$ of states Ψ_u and Ψ_w. To evaluate the coupling matrix element using DFT, we make a comparison with the matrix element H_{uw} from two nonorthogonal wave functions determined using the MOVB theory and in other calculations.[3,13,14,21,55] First, we define the one-particle exchange (or transition) density matrix and the exchange (transition) electron density as follows

$$\mathbf{D}_{uw} = \mathbf{C}^w(\mathbf{S}^{uw})^{-1}(\mathbf{C}^u)^{\mathrm{T}} = \mathbf{C}^w[(\mathbf{C}^u)^{\mathrm{T}} \mathbf{R} \mathbf{C}^w]^{-1}(\mathbf{C}^u)^{\mathrm{T}} \tag{39}$$

$$\rho_{uw}(\mathbf{r}) = \sum_{\mu\nu}^{m} |\chi_\mu(\mathbf{r})> (\mathbf{D}_{uw})_{\mu\nu} <\chi_\nu(\mathbf{r})| = \chi(\mathbf{r}) \mathbf{D}_{uw} \chi^{\mathrm{T}}(\mathbf{r}) \tag{40}$$

Clearly, the density matrix $\mathbf{D}_{uw}$ is of rank N (i.e., $\int \rho_{uw}(\mathbf{r}) d\mathbf{r} = N$; if $\mathbf{S}^{uw}$ is also of rank N), but it is not symmetric. Importantly, $\mathbf{D}_{uw}$ satisfies the general idempotency condition

$$\mathbf{D}_{uw} \mathbf{R} \mathbf{D}_{uw} = \mathbf{D}_{uw} \tag{41}$$

Consequently, the coupling matrix element between the two determinant states Ψ_u and Ψ_w can be computed in exactly the same way as that of Hartree−Fock energy expression; one simply replaces the symmetric ground-state one-particle density matrix with the asymmetric one-particle exchange density:

$$H_{uw}^{\mathrm{KS}} = S_{uw}^{\mathrm{KS}}\left\{ Tr[(\mathbf{D}_{uw})^{\mathrm{T}}\mathbf{h}] + \frac{1}{2}Tr[(\mathbf{D}_{uw})^{\mathrm{T}}\mathbf{J}\mathbf{D}_{uw}] \right. \\ \left. - \frac{1}{4}Tr[(\mathbf{D}_{uw})^{\mathrm{T}}\mathbf{K}\mathbf{D}_{uw}] + E_{\mathrm{nuc}}\right\} \tag{42}$$

Here, we have used the superscript KS to indicate that the exchange, or coupling matrix element, is computed using KS orbitals along with the pure Hartree−Fock exchange, $\mathbf{K}$.

Comparison with the KS-DFT energy expression (eq 13) shows that the only difference is in the exchange term. If we replace, as in ground-state KS-DFT calculations, the Hartree−Fock exchange term with the DFT exchange-correlation energy functional using the exchange density defined in eq 41 we may compute the VBDFT coupling matrix element between states characterized by the block-localized densities of $\rho_u(\mathbf{r})$ and $\rho_w(\mathbf{r})$ by

$$H_{uw}^{\mathrm{VBDFT}} \equiv E[\rho_{uw}(\mathbf{r})] = S_{uw}^{\mathrm{KS}}\left\{ Tr[(\mathbf{D}_{uw})^{\mathrm{T}}\mathbf{h}] + \frac{1}{2}Tr[(\mathbf{D}_{uw})^{\mathrm{T}}\mathbf{J}\mathbf{D}_{uw}] \right\} \\ + S_{uw}^{\mathrm{KS}}E_{xc}[\rho_{uw}(\mathbf{r})] + S_{uw}^{\mathrm{KS}}E_{\mathrm{nuc}} \tag{43}$$

Here, we have assumed that the approximate energy functional for the ground state is also applicable to the coupling energy between two interacting states using the transition density defined in eq 41. We note that it is possible to define the one-particle exchange density differently if a different form of the projection operator (eq 14) is used,[37,53] which corresponds to a different definition of the diabatic states. The projection operator in eq 14 satisfies Hermiticity.[34]

The adiabatic ground-state energy can be computed by a two-state, mixed valence bond and DFT model, that is, VBDFT, in which 1 specifies the reactant and 2 the product state, through the generalized secular equation:

$$\begin{vmatrix} E_{uu}^{\mathrm{BL\text{-}DFT}}[\rho_u(\mathbf{r})] - \varepsilon(\mathbf{X}) & E_{uw}^{\mathrm{BL\text{-}DFT}}[\rho_{uw}(\mathbf{r})] - \varepsilon(\mathbf{X})S_{uw}^{\mathrm{KS}} \\ E_{wu}^{\mathrm{BL\text{-}DFT}}[\rho_{wu}(\mathbf{r})] - \varepsilon(\mathbf{X})S_{wu}^{\mathrm{KS}} & E_{ww}^{\mathrm{BL\text{-}DFT}}[\rho_w(\mathbf{r})] - \varepsilon(\mathbf{X}) \end{vmatrix} = 0 \tag{44}$$

Here, we have used the superscripts to indicate the method by which the energy is computed. The numerical performance of eq 43 will be thoroughly addressed in a subsequent publication. In the present work, we describe an alternative, approximate method below, which will be applied to the test cases by using the off-diagonal matrix elements computed using the KS determinants.

In the second approach, the off-diagonal matrix element is approximated by one determined with the KS determinants of block-localized valence bond states. To make the electronic energy compatible with that computed using density functional theory, we assume that the relative energies computed at the DFT level and the Hartree−Fock level with KS determinants can be shifted by an amount corresponding to the absolute energy difference in the two diabatic states (a constant at a given geometry). Thus,

$$H_{uw} \approx H_{uw}^{\mathrm{KS}} + \frac{1}{2}S_{uw}(E_u[\rho_u(\mathbf{r})] + E_w[\rho_w(\mathbf{r})] - E_u^{\mathrm{KS}} - E_w^{\mathrm{KS}}) \tag{45}$$

where the BLDFT energies $E_u[\rho_u(\mathbf{r})]$ and $E_w[\rho_w(\mathbf{r})]$ for the states characterized by the densities $\rho_u(\mathbf{r})$ and $\rho_w(\mathbf{r})$ are determined by eq 13, and $E_u^{\mathrm{KS}}(\mathbf{K})$ and $E_w^{\mathrm{KS}}(\mathbf{K})$ are the Hartree−Fock energies determined by using the BLKS orbitals. As we shall see below, the use of the approximate coupling-matrix element of eq 45 can overcome the self-interaction errors in approximate functionals, but it typically yields barrier heights greater than the corresponding accurate results if only two states are used. Inclusion of the ionic configuration in a three-state model for the S_N2 reactions is examined here; the computed energies of reaction are found to be in accord with accurate results (below). We note that Wu et al. proposed a different way of estimating the off-diagonal matrix element,[49,56] which also involves the overlap scaled sum of the energies of the two diabatic states, but it is corrected by the overlap scaled constraint integrals. Equation 45 implicitly assumes that the relative change of

BLDFT and KS energies for the diabatic states is approximately similar.

Obviously, the method presented here is not restricted to two states, and it can easily be generalized to any number of states, defined by a given characteristic feature of charge and spin. In the present work, we also examined a three-state treatment of the S_N2 reactions. At this point, we note the difference of the two acronyms, BLDFT and VBDFT (or equivalently MSDFT). The multistate VBDFT Hamiltonian that we present here employs the BLDFT method to determine matrix elements; as noted in the beginning of this work, it can also be called a multistate density functional theory based on multiple Kohn−Sham reference densities of valence bond-like states. BLDFT is also a general theory within its own right, which can be used to investigate properties of the chemical bond,[45] such as resonance delocalization energy,[32,45] aromaticity,[57,58] polarization and charge transfer energy,[59−63] hyperconjugation and negative hyperconjugation,[32,64] and steric effects.[65,66] The VBDFT method is designed for the representation of the PES of chemical reactions, both in the gas phase and in solution and enzymes (simply by including the electrostatic potential in the term $v_{ext}(r)$ in eq 17). The diabatic states defined in eqs 33 and 34 can be used to define the Marcus reaction coordinate,[67] and to represent the solvent reaction coordinate along with the use of the adiabatic potential surface from the same or a different level of theory in condensed phase reactions. Methods for carrying out this type of calculation have been described previously.[13−15]

3. Computational Details

All calculations are carried out using the Xiamen Valence Bond[41] program and a modified version of GAMESS.[68] Geometry optimizations were performed with Gaussian 03.[69] The 6-31G+(d) basis set was used throughout for all calculations, except that for the H_2^+ system, which employs the aug-cc-pVTZ basis set. Geometries for the S_N2 reactions of [Cl^- + CH_3Cl], [F^- + CH_3Cl], and [HO^- + CH_3F] along the reaction coordinate defined below are optimized at the B3LYP/6-31+G(d) level. We used a version of the B3LYP implementation that employs the VWN5 functional in GAMESS, which differs from the standard B3LYP employing the VWN1 functional.[70−73] All results are obtained using the VBDFT method employing eq 45 in the present study.

To describe the change in energy and wave function of the two Lewis bond states for each reaction, we define the reaction coordinate as the difference between the bond length of the central carbon and the leaving group $R(C−L)$, where L = Cl, Cl, and F, and that of the nucleophile and the central carbon $R(Nu−C)$, where Nu = Cl, F, and O, respectively, for the three reactions:

$$Rc = R(C−L) − R(Nu−C) \qquad (46)$$

Of course, one can use other definitions to monitor the progress of the reaction, including the difference between the corresponding bond orders or energies of the two Lewis bond states. The geometrical variable, corresponding to the asymmetric bond stretch coordinate, is a good choice and chemically intuitive for the S_N2 reactions.
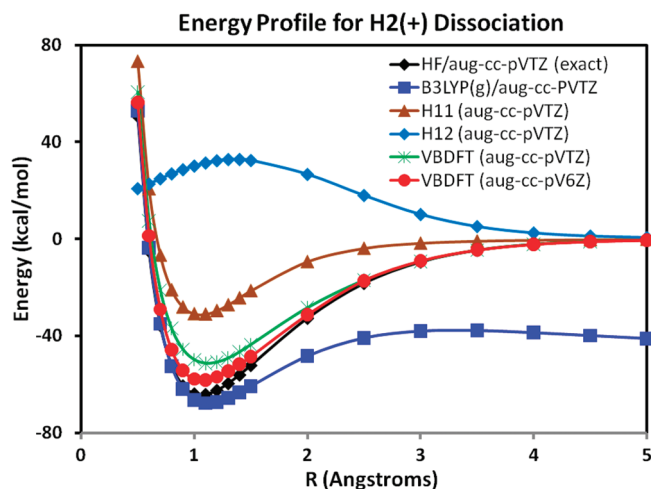
Figure 1. Computed adiabatic ground-state energy profile for the diassociation of H_2^+ using Hartree−Fock (HF) theory (black), density functional theory with the hybrid B3LYP functional (blue), and multireference valence bond-density functional theory (VBDFT) using the B3LYP functional with the aug-cc-pVTZ basis set (green) and with the aug-cc-pV6Z basis set (red). The diabatic energy profile for the degenerate block-localized electronic configuration $\Psi([H^+][H\bullet])$ computed using the BLDFT method is shown in brown, and the coupling energy V_{12} is given in light blue, both using the aug-cc-pVTZ basis functions.

4. Results and Discussion

We first examine the simple but illustrative case of the dissociation of H_2^+ to show that the BLDFT method can effectively localize charge configurations and the localized states can be used in VB-like configuration interaction calculations to yield a qualitatively correct energy profile. Then, we present the results for the nucleophilic substitution reactions of [Cl^- + CH_3Cl], [F^- + CH_3Cl], and [HO^- + CH_3F] that are included in the DBH24 database designed for testing the performance of new functionals.

A. Energy Profile of H_2^+. Recently, Yang and co-workers[74] vividly illustrated the well-known self-interaction error (charge delocalization error) in current approximate functionals by considering the simple one-electron system H_2^+. In this case, the Hartree−Fock result is exact with a given basis set, and DFT describes the chemical bond very well. However, approximate functionals, as illustrated with B3LYP in Figure 1, fail badly as the distance between the two protons increases;[74] the potential energy profile first reaches a barrier of about 20 kcal/mol above the minimum and then falls off continuously because the electron is delocalized over two centers at infinite separation. This has been attributed to the self-interaction error.[74,75] This error can also be regarded as a need for multiconfigurational treatment of the dissociation process (also in the case of H_2 associated with spin delocalization). Cohen et al. pointed out that this charge delocalization error causes the computed barrier for a chemical reaction to be too low, a molecular polarizability that is too high, and unphysical charge transfer in molecular electronic devices using approximate functionals.[74]

In VBDFT, we employ two degenerate block-localized configurations, corresponding to the electron localized on

BLDFT and Diabatic Coupling

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2709**

the "left" proton, $\Psi_1([H\bullet][H^+])$, and on the "right" proton, $\Psi_2([H^+][H\bullet])$, respectively. The energy profile of the block-localized VB state is shown in Figure 1 by the curve denoted as H_{11}, which gives the electrostatic and polarization interaction energy between a localized (but polarizable) hydrogen atom and a proton. The computed interaction energy at the minimum is -31 kcal/mol, slightly less than half of the total binding energy at the Hartree−Fock level. Using the wavefunction-based estimate of the off-diagonal element in the VBDFT method, we obtain a bond-dissociation energy about 10 kcal/mol smaller than the exact result. Although both Hartree−Fock and B3LYP results are well converged with the aug-cc-pVTZ basis set, the use of a larger basis set in VBDFT calculations can further improve the estimated binding energy. For example, using the aug-cc-pV6Z basis, the calculated VBBLD result is -58.3 kcal/mol compared with the Hartree−Fock (exact) value of -63.5 kcal/mol. Thus, the block-localized density functional theory is more sensitive to the atomic orbital basis functions because of the extreme localization of Kohn−Sham orbitals. Importantly, the qualitative behavior of the dissociation curve at large interatomic distances is correctly reproduced. Wu and co-workers showed that their CDFT-CI approach can also effectively describe the correct dissociation energy.[49]

Depicted in Figure 1 also is the coupling energy, defined by eq 3 ($V_{12} = |H_{12} - \varepsilon_g S_{12}^{KS}|$). The largest coupling energy is in the bonding region and goes to zero as the interatomic separation becomes large. Figure 1 shows that the BLDFT method can effectively localize charge density for VB-like configurations. It is interesting to note that the CDFT method used by Wu and co-workers employs a spatial integration to constrain electron density, which cannot be exactly unity on the constrained atom unless the integration is over the entire space.[49] This limitation is not obvious for molecular systems with more than one electron, but its implication, due to the use of fully delocalized Kohn−Sham orbitals, is clearly illustrated in the one-electron case. In the BLDFT method, integration of the electron density defined by eq 12 yields exactly the number of electrons in the localized fragment (one for a hydrogen atom).

B. S_N2 Reactions. *B.1. Diabatic Energies.* We consider the diabatic potential energy profiles for the three model S_N2 reactions, [$Cl^- + CH_3Cl$], [$F^- + CH_3Cl$], and [$HO^- + CH_3F$], included in the DBH24 (24 diverse barrier heights) database.[76] In each case, the reactant state is defined by the VB configuration characterized by the substrate Lewis structure in the presence of the nucleophile as a "spectator ion" (eq 33) and the product state by the product Lewis structure in the presence of the leaving group ion (eq 34).[3] In addition, we constructed a three-state model by including the ionic configuration:

$$\Psi_x(\mathbf{X}) = N_x \hat{A}\{\Omega^6[Nu^-]\,\Omega^7[CH_3^+]\,\Omega^8[L^-]\} \qquad (47)$$

The VB configuration of eq 47 closely resembles the structure of the transition state. The reactant and product diabatic energy profiles for the three S_N2 reactions, computed using the BLDFT method and Hartree−Fock theory with BLKS orbitals, are shown in Figure 2, along with the VBDFT
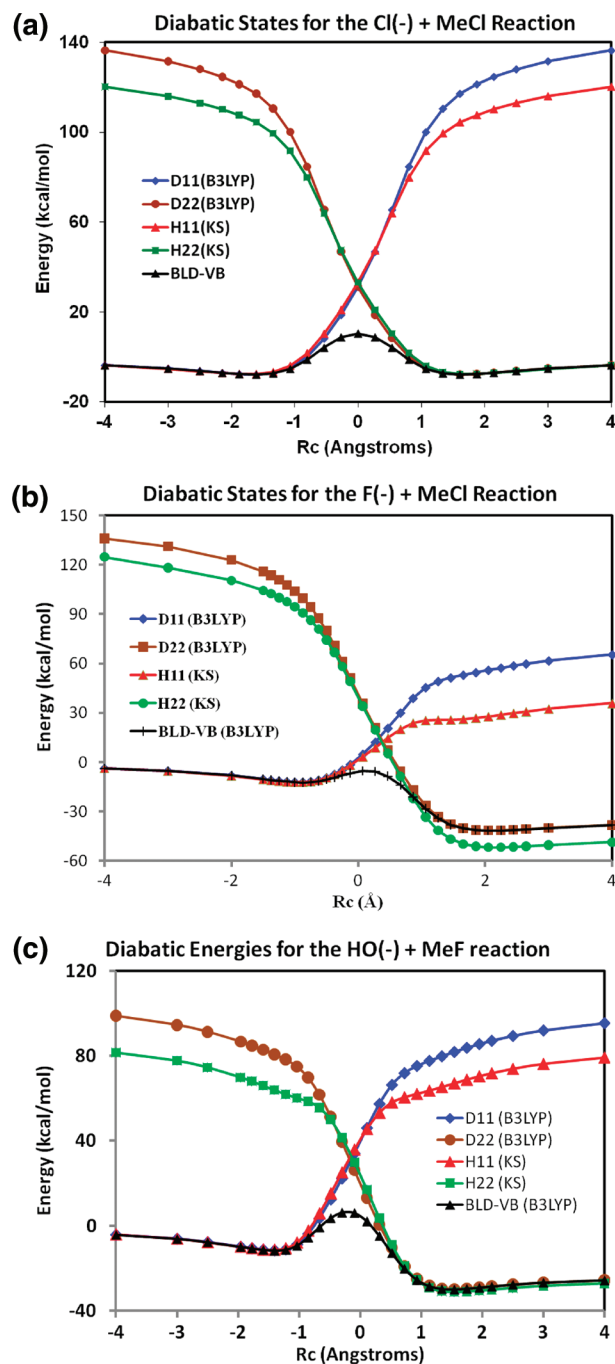
Figure 2. Diabatic energy profiles of the reactant state computed using the BLDFT method (D11 in blue) and wave function theory with the block-localized Kohn−Sham orbitals (H11(KS) in red) and of the product state using the BLDFT method (D22 in brown) and wave function theory (H22(KS) in green) for (a) [$Cl^- + CH_3Cl$], (b) [$F^- + CH_3Cl$], and (c) [$HO^- + CH_3F$] reactions.

adiabatic ground-state potential energy. The relative energies are determined with respect to the infinitely separated nucleophile and substrate species (reactant state). Thus, the absolute electronic energies of the diabatic and adiabatic states can be obtained by adding the sum of the corresponding DFT energies of the separate molecules.

The three S_N2 reactions represent very different features in the electron-withdrawing power of the nucleophiles and the leaving groups. The chloride ion exchange reaction is

Table 1. Computed Binding Energies (kcal/mol) for the Formation of the Ion-Dipole (ΔE_1) Complex, the Barrier Height Relative to the the IP Complex ($\Delta E^{\ddagger}$), the Relative Energy between the Reactant and Product Ion−Dipole Complexes (ΔE_2), and the Net Energy of Reaction (ΔE_{rxn}) between the Product and Reactant States for the S$_N$2 Reactions (see Scheme 1)

	DFT-VB(2)[a]	DFT-VB(3)[a]	B3LYP[a]	M06-2X[b]	CCSD(T)[b]	NSBH6[c]
			[Cl⁻ + CH₃Cl]			
ΔE_1	−7.8	−9.0	−9.5	−11.3	−10.9	−10.5
$\Delta E^{\ddagger}$	18.3	12.5	8.6	13.2	12.6	13.6
ΔE_2	−7.8	−9.0	−9.5	−11.3	−10.9	−10.5
ΔE_{rxn}	0.0	0.0	0.0	0.0	0.0	0
			[F⁻ + CH₃Cl]			
ΔE_1	−12.3	−15.0	−16.0	−17.1	−16.1	−16.4
$\Delta E^{\ddagger}$	7.1	2.5	0.3	3.6	2.9	2.9
ΔE_2	−7.4	−8.0	−8.4	−10.0	−9.7	−10.5
ΔE_{rxn}	−34.3	−34.3	−34.3	−35.8	−32.2	−32.7
			[HO⁻ + CH₃F]			
ΔE_1	−11.6	−10.0	−13.7	−14.6	−13.8	−13.7
$\Delta E^{\ddagger}$	17.9	10.7	6.1	12.3	10.8	11.0
ΔE_2	−6.6	−7.5	−8.1	−9.6	−9.2	
ΔE_{rxn}	−23.3	−23.3	−23.3	−20.7	−19.6	−20.1

[a] The VWN5 functional was used in the B3LYP implementation in GAMESS with the 6-31+G(d) basis set. Values in parentheses are computed using three VB configurations. [b] Computed using the aug-cc-pVTZ basis set at the QCISD/MG3 geometry. [c] Ref 76.

thermal neutral, whereas as the reaction between F⁻ and CH₃Cl and that between HO⁻ and CH₃F are highly exothermic (Table 1). The diabatic energy differences between the reactant and product ion−dipole complexes are, respectively, 0.0, −29.5, and −18.4 kcal/mol for the three reactions from the BLDFT model, which are in good accord with the difference of the adiabatic ground state (Table 1). In the reactant and product state region up to the crossing point (transition state), the respective diabatic energies computed using the BLDFT method and block-localized MOs are in very good accord (Figure 2). However, beyond the crossing point, BLDFT results show a somewhat greater energy increase than that of the Hartree−Fock results (using BLKS orbitals, i.e., H_{ii}^{KS}) as the reaction coordinates further deviates from equilibrium geometries. Note that the latter is not used as the diabatic state energy in the VBDFT Hamiltonian, and it is shown to illustrate the trend of the energy change in comparison with the BLDFT results. Thus, the appearance that H_{22}^{KS} is below the VBDFT adiabatic ground-state energy for the [F⁻ + CH₃Cl] reaction in Figure 2b is due to different energy scales between the absolute energies from DFT and Hartree−Fock methods (in VBDFT, the variational principle ensures that $\varepsilon_g \leq H_{ii}^{BL\text{-}DFT}$).

Figure 3 shows the partial atomic charges of the leaving group in the reactant state configuration, and those of the nucleophiles in the product configuration as a function of Rc for all three reactions. The two-state model, VBDFT(2), where the number in parentheses specifies the number of VB states, was used in these analyses. By definition, the total charge densities of the nucleophiles in the reactant states and those of the leaving group in the product states are unity (eqs 33 and 34), which are confirmed by the corresponding Mulliken populations (not shown). Figure 3 highlights the effects of polarization of the Lewis structures by the neighboring ions in the respective diabatic states. Consequently, although the total charge of the subgroups of the substrate species in the reactant state and of the product Lewis structure in the product state are constrained to be zero by virtue of block-localization of the total electron density, the electronic structures of these subgroups are

Partial Atomic Charges
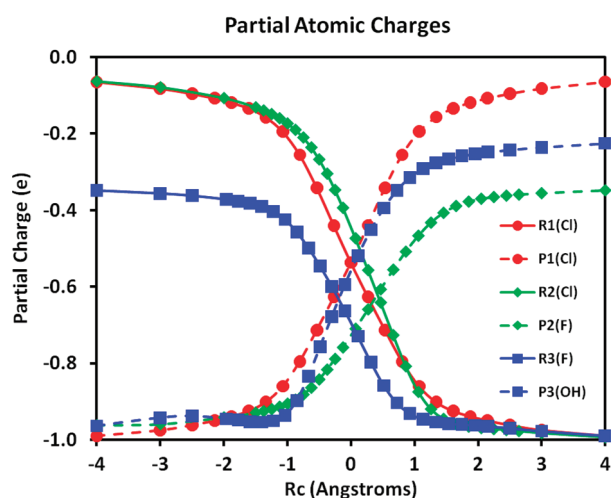
Figure 3. Mulliken population charges for the leaving group in the reactant diabatic state in the [Cl⁻ + CH₃Cl] reaction (red), R1(Cl); in the [F⁻ + CH₃Cl] reaction (yellow), R2(Cl); and in the [HO⁻ + CH₃F] reaction (blue), R3(F). Mulliken population charges for the nucleophile in the product diabatic state configuration are shown as dashed lines of the same colors for the three reactions given above. Partial charges are given in atomic units.

significantly polarized by the "neighboring" ions. For example, in the reaction of [F⁻ + CH₃Cl], the subgroup representing the reactant Lewis structure [CH₃ − Cl] ([CH₃ − L] in eq 33) is essentially ionized by the F⁻ ion at the product state geometry. In all three reactions, there is significant ionic character developed in the neutral subgroups both in the reactant and product diabatic states at the crossing point (transition state), suggesting that it may be useful to include the ionic VB configuration (eq 47) in the VBDFT Hamiltonian. It is also interesting that the CH₃Cl subgroup in the reactions of [Cl⁻ + CH₃Cl] and [F⁻ + CH₃Cl] exhibits different charge polarizations by the nucleophiles, although the trends of the partial atomic charge on the chlorine atom are similar.

B.2. Energy Profiles of the Adiabatic Ground State. The adiabatic ground-state potential energy profiles for the three

BLDFT and Diabatic Coupling

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2711**

(a)

(b)

(c)
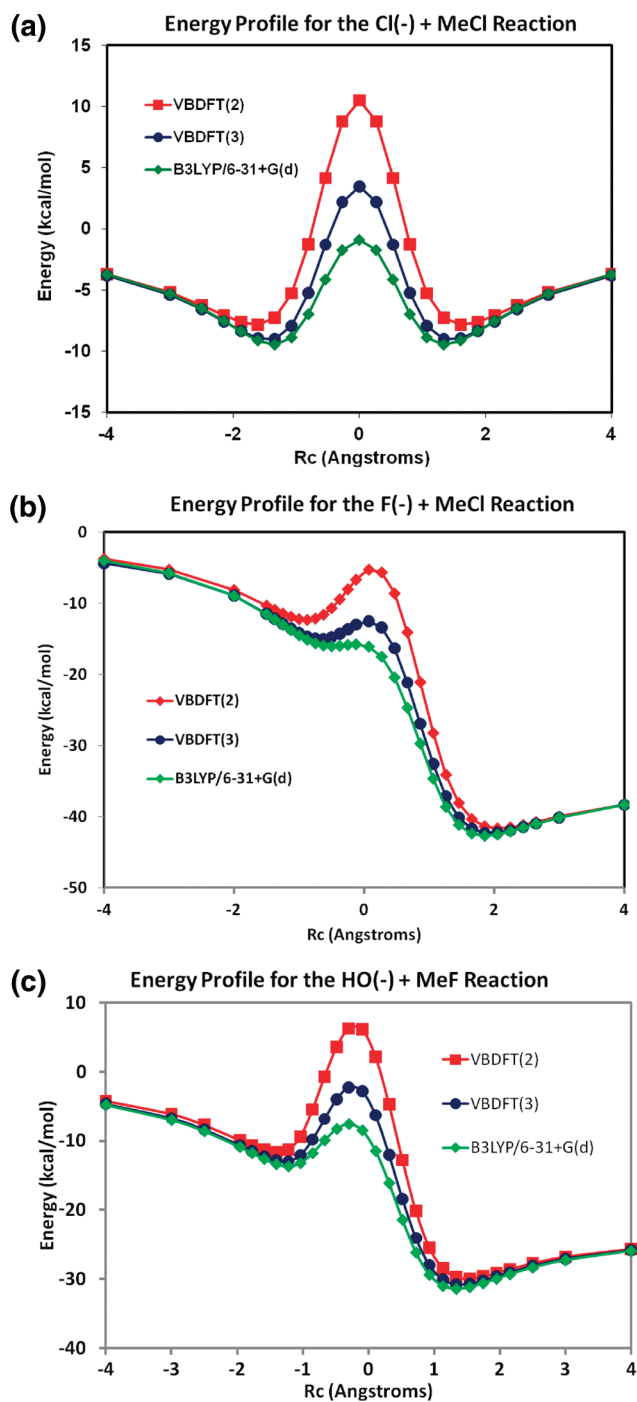
Figure 4. Adiabatic ground-state potential energy profiles determined using the hybrid B3LYP density function theory (green), the VBDFT(2) method with two states (red), and VBDFT(3) with three states (blue) for (a) the $[Cl^- + CH_3Cl]$, (b) the $[F^- + CH_3Cl]$, and (c) the $[HO^- + CH_3F]$ reactions.

S_N2 reactions considered in this work are given in Figure 4 along the reaction coordinate defined by the asymmetric stretch coordinate (eq 46). These effective diabatic states are not unique,[3,4,7,17−19] and in ab initio VB theory, this can be rationalized by an arbitrary (although quite reasonable) partition of the ionic structures (eq 47) into the two diabatic Lewis configurations.[3] In BLDFT, they are treated by the block-localized electronic structure and are fully delocalized within each subgroup due to the use of a single determinant that produces the exact electron density with the Kohn−Sham

approach. As noted previously, the reduction of VB configurations from a complete VB space to a two-state model leads to a loss of static correlation, in addition to reduced delocalization effects. There is no reason to expect a true ab initio two-state model to be able to reproduce the exact results (unless these two states are reconstructed from the full VB wave function as in the CDC-MOVB (consistent diabatic configuration) method described in ref 3 or transformed from the adiabatic ground state and its coupling excited state[3,4,28]). This is clearly reflected by the results shown in Figure 4, in which the computed barrier heights from the two-state VBDFT(2) model are noticeably greater than the adiabatic B3LYP (the same level of theoretical model) results.

For the three S_N2 reactions, the hybrid functional underestimates the barriers considerably, a well-known problem for these reactions, which may be attributed to the presence of two degenerate VB-like diabatic states at the transition state (Figure 2).[3] The relatively high barriers from VBDFT(2) calculations in comparison with accurate results[76] are due to three main factors. First, the block-localized Kohn−Sham orbitals are more sensitive to the basis set used than the standard ground-state calculations, as shown in the one-electron case (Figure 1). Thus, larger basis functions may improve the accuracy of the computed energies of the diabatic states as well as the off-diagonal matrix element. For the $[Cl^- + CH_3Cl]$ reaction, the barrier decreases by 0.2 kcal/mol on switching to the 6-311+G(d,p) basis set. In the case of the $[HO^- + CH_3F]$, the barrier is reduced by 1.0 kcal/mol using aug-cc-pVDZ and 1.4 kcal/mol using aug-cc-pVTZ. Second, the off-diagonal matrix elements are computed with an approximate method. Third, other valence bond configurations also contribute significantly to the stabilization of the transition state (see below). Nevertheless, a key finding of Figure 4 is that the present multistate VBDFT method can help correct the self-interaction errors in standard approximate functionals for group-transfer reactions examined here.

To illustrate the effects of the ionic configuration on the computed reaction barrier, we include a third state,[3,13] eq 47, in the VBDFT(3) calculation for the three S_N2 reactions. Remarkably, the inclusion of the ionic configuration significantly improves the computed barrier heights, in comparison with the M06-2X and CCSD(T) results.[76] The reaction involving $[HO^- + CH_3F]$ is a particularly challenging case in the DBH24 database because of the high electronegativity and small size of the nucleophile and leaving group ions.[76] As shown in Figure 3, the covalent bond for both the reactant and product diabatic states are significantly more polarized than the other two reactions.

Listed in Table 1 are the relative energies of all stationary points (depicted in Scheme 1) for the three S_N2 reactions along the reaction coordinate. The VBDFT(3) results are in good accord with currently the most accurate DFT (M06-2X/aug-cc-pVTZ) results,[76] as well as with high-level wave function theories (CCSD(T)/aug-cc-pVTZ),[76] suggesting that multistate VBDFT can provide an adequate description of the reaction PES. Finally, we note that the distances in the

Scheme 1. Definition of Relative Energies
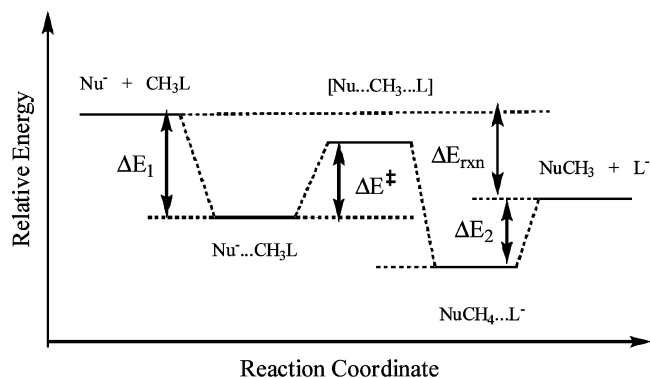
ion−dipole complex minima are slightly longer by about 0.1−0.2 Å from DFT-VB calculations than the B3LYP results.

B.3. Polarization, Charge Transfer, and Coupling Energies. The BLDFT method used to construct the diabatic state in the multireference VBDFT approach is a versatile theoretical model that can be used to study chemical bonding properties and to decompose the component of intermolecular interaction energies.[45] It provides a useful tool to analyze DFT results to gain chemical insight into the system. In the Appendix, we illustrate the utility of the BLDFT method and shed light on the contributing components of the interaction energies for the ion−dipole complexes.

It is useful to consider the energy difference between the VBDFT (ε_g^{VBDFT}) and the conventional DFT ($E^{DFT}[\rho(Nu^-\cdot CH_3L)]$) method:

$$\Delta E_{rcr} = \varepsilon_g^{VBDFT} - E^{DFT}[\rho(Nu^-\cdot CH_3L)] \qquad (48)$$

where the superscript DFT in $E^{DFT}[\rho(Nu^-\cdot CH_3L)]$ is used to emphasize that the energy is determined by conventional KS-DFT. In VBDFT, each diabatic state is completely defined by its Hamiltonian, and the Hohenberg−Kohn theorem ensures that its ground-state energy is solely determined by the block-localized density.[78,79] Obviously, localization of the electron density loses electronic delocalization (resonance) energy between different fragments or blocks (thus, the steep rise of diabatic energy shown in Figure 2). In addition, dynamic correlation effects will also be different in the diabatic states from that of the adiabatic ground state. VB configuration interaction in the complete VB space recovers delocalization effects and introduces static correlation; however, resonance and dynamic correlation effects are not necessarily fully restored in the adiabatic ground state if a minimal number of VB configurations is used. ΔE_{rcr} is a measure of the residual correlation−resonance (rcr) effect in the multistate, multireference VBDFT method for stable molecules where one diabatic state dominates the VB wave function. If Hartree−Fock theory is used, the dominant effect in ΔE_{rcr} is the residual resonance energy, thereby providing an estimate of residual correlation effects.

At the transition state (diabatic crossing point), the reactant and product diabatic configurations are degenerate, which has the largest self-interaction error along the reaction path. The quantity ΔE_{rcr} now gives a reasonable indication of the

Table 2. Computed Energy Components and Coupling Effects for Ion−Dipole Complexes[a]

complex	ΔE_{int}	ΔE_R	ΔE_{stat}	ΔE_{pol}	ΔE_{CT}	$\Delta E_{rcr}(2)$	$\Delta E_{rcr}(3)$
$Cl^-\cdot CH_3Cl$	−9.5	0.5	−5.0	−2.7	−2.3	2.2	0.5
$Cl^-\cdot CH_3F$	−8.6	0.2	−5.7	−1.7	−1.5	1.4	0.4
$F^-\cdot CH_3Cl$	−16.0	2.7	−4.7	−7.4	−6.7	5.4	1.3
$F^-\cdot CH_3OH$	−8.0	0.6	−3.2	−3.7	−1.8	1.7	0.6
$HO^-\cdot CH_3F$	−13.7	0.9	−8.0	−3.9	−2.7	2.4	0.7

[a] All energies are given in kilocalories per mole.

Table 3. Computed Coupling Energy V_{12} and Residue Correlation−Resonance ΔE_{rcr} Energy (kcal/mol) at the Transition State for the Nucleophilic Substitution Reactions

	$\Delta E_{rcr}(2)^a$	$\Delta E_{rcr}(3)^a$	V_{12}	$\Delta(\Delta E_{rcr})^b$
$[Cl^- + CH_3Cl]$	11.4	4.4	20.7	7.0
$[F^- + CH_3Cl]$	10.5	3.2	20.2	7.3
$[HO^- + CH_3F]$	13.8	5.3	23.2	8.5

[a] The number in parentheses indicates the number of configurations used in VBDFT calculations. [b] $\Delta(\Delta E_{rcr}) = \Delta E_{rcr}(2) - \Delta E_{rcr}(3)$.

relative magnitude of the self-interaction energy error, keeping in mind that ΔE_{rcr} also contains delocalization energy not fully recovered by the multistate VB model. Of course, in such a multiconfigurational treatment, there is also a concern of double counting of correlation effects,[80] which is an issue to be investigated in the future by comparison with experimental results.

Table 2 shows that, in the present VBDFT method, the ΔE_{rcr} term is found to be positive, indicating that the loss in charge delocalization and dynamic correlation in constructing the diabatic states is not fully compensated for by the multiconfigurational VB Hamiltonian. Not surprisingly, the absolute value of the ΔE_{rcr} energy is similar to that of the charge transfer component. This is consistent with the interpretation that the ΔE_{rcr} term is mainly due to reduction in the resonance delocalization and dynamic correlation effect in the ion−dipole complex state. It is interesting to notice that inclusion of the ionic configurations significantly reduces ΔE_{rcr} values, indicating the importance of the ionic configuration even at the stable, ion−dipole complex geometry.

The computed coupling energies, $V_{12} = |H_{12} - \varepsilon_g S_{12}|$, using VBDFT(2) at the transition state are shown in Table 3, and the variation of V_{12} along the reaction coordinate is shown in Figure 5. For all three reactions, V_{12} has the largest values at the transition state and decays rapidly as the reaction coordinate moves away from the transition state. V_{12} is essentially zero at the ion−dipole complex structure. For a comparison with ion−dipole complexes above, there is greater reduction in ΔE_{rcr} with the inclusion of the ionic configuration at the transition state (Table 3), indicating that the ionic configuration makes greater contributions than the stabilization of the reactant state energy. It is interesting to note that the change, $\Delta(\Delta E_{rcr}) = \Delta E_{rcr}(2) - \Delta E_{rcr}(3)$, is nearly the amount of energy needed to increase V_{12} in the two-state model to yield the correct barrier height.

5. Conclusions

In this article, we describe BLDFT for the construction of valence-bond-like diabatic electronic states, suitable for the

(a) Coupling Energy for the Cl(-) + MeCl Reaction

(b) Coupling Energy for the F(-) + MeCl Reaction

(c) Coupling Energy for the HO(-) + MeF Reaction
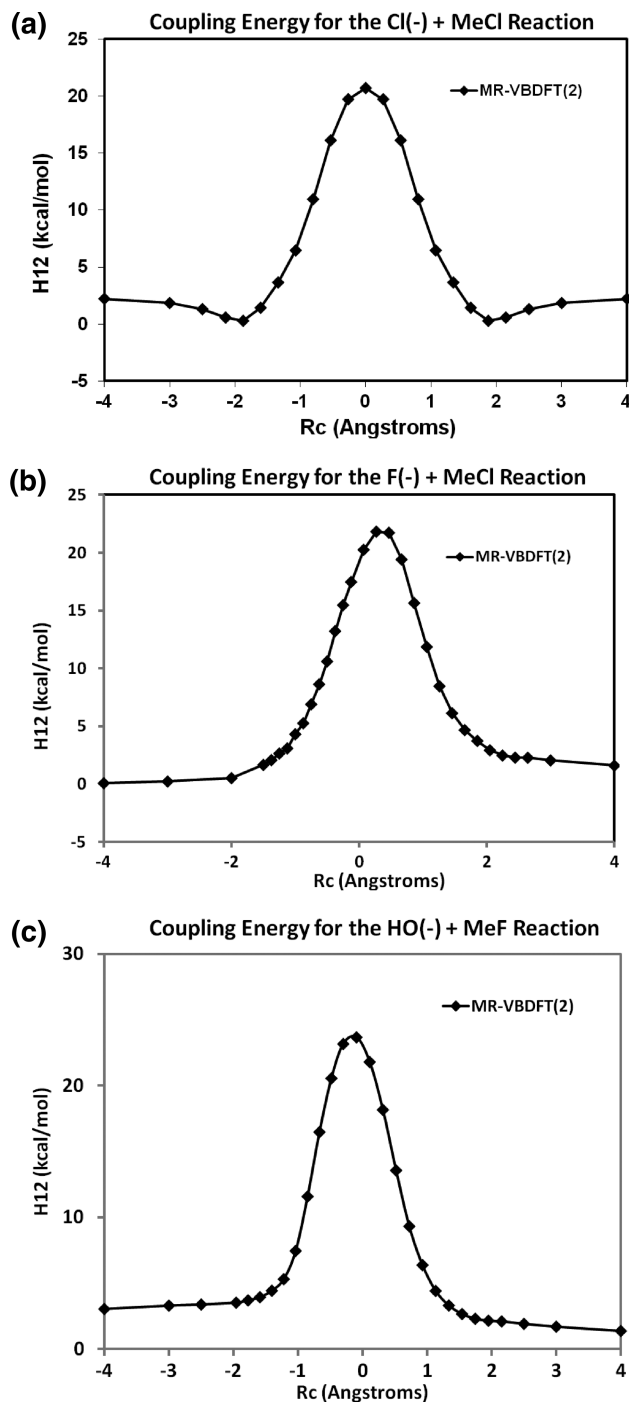
Figure 5. Computed coupling energy, V_{12}, for (a) the $[Cl^- + CH_3Cl]$, (b) the $[F^- + CH_3Cl]$, and (c) the $[HO^- + CH_3F]$ reactions.

study of electron transfer reactions and for the representation of the reactive potential energy surface. The BLDFT method has been used previously in intermolecular interaction energy decomposition analysis, which is an extension of the corresponding wave function theory approach. Here, we further examine some of the properties of the BLDFT method. Importantly, the BLDFT method is used to formulate a multistate density functional theory in the framework of the valence bond, making use of a reduced configuration representation and of a single determinant strategy for each VB configuration. Thus, the method is equivalent to a valence bond Hamiltonian using density functional theory, which can

be called the VBDFT method. The most important result of this study is to propose a functional estimate of the off-diagonal matrix elements of the VB Hamiltonian, assuming that the overlap integral between KS determinants of nonorthogonal block-localized orbitals is a reasonable approximation to the overlap of the diabatic state defined by their ground-state electron density, and the exchange-correlation functional for the ground state can also be used to describe the exchange (transition) matrix elements.

We also described an approximate approach in which the off-diagonal matrix element is computed by using the determinant of nonorthogonal BLKS orbitals. The use of the approximate algorithm in VBDFT calculations was illustrated by applications to the bond dissociation of H_2^+, a simple example to illustrate the self-interaction energy error in approximate functionals, and a set of three nucleophilic substitution reactions in the DBH24 database designed for validating computational accuracy. In the dissociation of H_2^+, the DFT-VB method yields the correct asymptotic behavior as the two protons are stretched to infinity, whereas approximate density functionals fail badly. The computed bonding energy from the VBDFT method is more dependent on the size of basis set than that of the hybrid B3LYP functional due to the extreme localization of the Kohn–Sham orbitals. For the S_N2 nucleophilic substitution reactions, the hybrid functional B3LYP severely underestimates the barrier heights, while the two-state VBDFT model overestimates the barrier heights, in comparison with results from the M06-2X functional and CCSD(T) wave function model. With the inclusion of the ionic configuration, the VBDFT results are found to be in good accord with these high-level accurate results.

The BLDFT method is a versatile theory that can be used to analyze conventional DFT results to gain insight into chemical bonding properties, and it is illustrated by examining the intricate energy contributions in the ion–dipole complex stabilization in the Appendix. It is found that electrostatic interactions provide the largest contribution to the binding interaction in the ion–dipole complex, followed by polarization effects and charge transfer contributions. At the transition state, diabatic coupling is very strong, lowering the diabatic state energy by 26 to 32 kcal/mol to stabilize the transition state of the adiabatic ground-state potential surface. The overstabilization of the transition state energy using approximate density functionals because of the presence of degenerate diabatic states is corrected with the use of the multistate, multireference VBDFT Hamiltonian. It is of interest to use the BLDFT and VBDFT methods to study chemical and electron reactions in solution and in enzymes.

Appendix: Energy Decomposition Analysis of Intermolecular Interactions

The interaction energy of an ion−dipole complex is computed by

$$\Delta E_{int} = \Delta E_1 = E[Nu^- \cdot CH_3L] - (E[Nu^-] + E[CH_3L]) \quad (A1)$$

where $E[Nu^- \cdot CH_3L]$, $E[Nu^-]$, and $E[CH_3L]$ are, respectively, the energies of the complex, individual ion, and molecule. ΔE_{int} can be decomposed into the following energy terms, including geometry distortion ΔE_R, static electrostatic interaction energy ΔE_{stat}, electronic polarization ΔE_{pol}, and charge transfer energy ΔE_{CT}:[32,33,37,38,45,59,60,62−66,77]

$$\Delta E_1 = \Delta E_R + \Delta E_{stat} + \Delta E_{pol} + \Delta E_{CT} \quad (A2)$$

ΔE_R represents the energy penalty due to geometry distortion of the individual molecules in the complex, which is a small factor for the ion−dipole complex (Table 2).

The remaining terms are determined using the optimized ion−dipole complex geometry. ΔE_{stat}, which is the sum of Coulomb and exchange repulsion energies, is determined from the energy of the block-localized electronic configuration with the subgroup KS orbitals taken from the isolated fragments by

$$\Delta E_{stat} = E[\rho_o^{BL\text{-}DFT}(Nu^- \cdot CH_3L)] - E[\rho_o(Nu^-)] - E[\rho_o(CH_3L)] \quad (A3)$$

where the subscript "o" indicates that the electron density is computed using the antisymmetric wave function constructed using the KS orbitals of the individual (isolated) molecules, in parentheses, optimized when they are infinitely separated. The energy $E[\rho_o^{BL\text{-}DFT}(Nu^- \cdot CH_3L)]$ is that of the first iteration in BLKS-SCF calculations when the optimized KS orbitals for Nu^- and CH_3L are read in as the initial guess.

Relaxation of the subgroup BLKS orbitals in the presence of the other subgroups yields electronic polarization energy in the complex:

$$\Delta E_{pol} = E[\rho^{BL\text{-}DFT}(Nu^- \cdot CH_3L)] - E[\rho_o^{BL\text{-}DFT}(Nu^- \cdot CH_3L)] \quad (A4)$$

where $\rho^{BL\text{-}DFT}(Nu^- \cdot CH_3L)$ is the fully relaxed (optimized) block-localized density. The extension of the basis set expansion to the full molecular system, that is, conventional DFT calculations, results in the energy component due to charge transfer:

$$\Delta E_{CT} = E[\rho(Nu^- \cdot CH_3L)] - E[\rho^{BL\text{-}DFT}(Nu^- \cdot CH_3L)] \quad (A5)$$

Notice that, for the reactant diabatic state of the S_N2 reactions considered here, $E[\rho^{BL\text{-}DFT}(Nu^- \cdot CH_3L)] = H_{11}$, whereas for the product complex, $E[\rho^{BL\text{-}DFT}(Nu^- \cdot CH_3L)] = H_{22}$.

Table 2 lists the energy components for the five unique ion−dipole complexes in the three nucleophilic substitution reactions. We do not perform energy decomposition analysis for the transition state because the isolated subgroups lose

meaning in such highly distorted geometries. In all cases, except the complex $F^- \cdot CH_3Cl$, the formation of the complex results in relatively small geometrical distortion (ΔE_R). In the $F^- \cdot CH_3Cl$ structure, there are also relatively greater effects from polarization and charge transfer than for the other four complexes, suggesting significant covalent bond character has been formed. Overall, static−electrostatic interaction has the greatest contribution, followed by polarization and charge transfer, to the stabilization of ion−dipole complexes.

References

(1) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265−322.

(2) Subotnik, J. E.; Yeganeh, S.; Cave, R. J.; Ratner, M. A. *J. Chem. Phys.* **2008**, *129*, 244101.

(3) Song, L.; Gao, J. *J. Phys. Chem. A* **2008**, *112*, 12925−12935.

(4) Valero, R.; Song, L.; Gao, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1−22, Erratum: **2009**, *5*, 2191.

(5) Mo, Y. *J. Chem. Phys.* **2007**, *126*, 224104.

(6) Newton, M. D. *Chem. Rev.* **1991**, *91*, 767−792.

(7) Shaik, S.; Hiberty, P. C. *A Chemist's Guide to Valence Bond Theory*; John Wiley & Sons, Inc.: Hoboken, NJ, 2008; p 1.

(8) Chang, Y. T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884−5888.

(9) Maupin, C. M.; Wong, K. F.; Soudackov, A. V.; Kim, S.; Voth, G. A. *J. Phys. Chem. A* **2006**, *110*, 631−639.

(10) Schlegel, H. B.; Sonnenberg, J. L. *J. Chem. Theory Comput.* **2006**, *2*, 905−911.

(11) Kim, Y.; Corchado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718−2735.

(12) Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13530−13536.

(13) Mo, Y.; Gao, J. *J. Comput. Chem.* **2000**, *21*, 1458−1469.

(14) Mo, Y.; Gao, J. *J. Phys. Chem. A* **2000**, *104*, 3012−3020.

(15) Gao, J.; Garcia-Viloca, M.; Poulsen, T. D.; Mo, Y. *Adv. Phys. Org. Chem.* **2003**, *38*, 161−181.

(16) Song, L.; Mo, Y.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 174−185.

(17) Mead, C. A.; Truhlar, D. G. *J. Chem. Phys.* **1982**, *77*, 6090−6098.

(18) Pacher, T.; Cederbaum, L. S.; Koppel, H. *J. Chem. Phys.* **1988**, *89*, 7367.

(19) Sidis, V. *Adv. Chem. Phys.* **1992**, *82*, 73.

(20) Baer, M. *Beyond Born-Oppenheimer: Electronic Nonadiabatic Coupling Terms and Conical Intersections*; Wiley: New York, 2006; p 1.

(21) Farazdel, A.; Dupuis, M.; Clementi, E.; Aviram, A. *J. Am. Chem. Soc.* **1990**, *112*, 4206−4214.

(22) Lu, D.; Chen, G.; Perry, J. W.; Goddard, W. A., III. *J. Am. Chem. Soc.* **1994**, *116*, 10679−10685.

(23) Thompson, W. H.; Blanchard-Desce, M.; Hynes, J. T. *J. Phys. Chem. A* **1998**, *102*, 7712−7722.

(24) Thompson, W. H.; Blanchard-Desce, M.; Alain, V.; Muller, J.; Fort, A.; Barzoukas, M.; Hynes, J. T. *J. Phys. Chem. A* **1999**, *103*, 3766−3771.

BLDFT and Diabatic Coupling

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2715**

(25) Voityuk, A. A.; Rosch, N. *J. Chem. Phys.* **2002**, *117*, 5607–5616.

(26) Prytkova, T. R.; Kurnikov, I. V.; Beratan, D. N. *J. Phys. Chem. B* **2005**, *109*, 1618–1625.

(27) Cave, R. J.; Newton, M. D. *Chem. Phys. Lett.* **1996**, *249*, 15–19.

(28) Prezhdo, O. V.; Kindt, J. T.; Tully, J. C. *J. Chem. Phys.* **1999**, *111*, 7818–7827.

(29) Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *125*, 164105.

(30) Wu, Q.; Van Voorhis, T. *J. Chem. Theory Comput.* **2006**, *2*, 765–774.

(31) Mo, Y.; Peyerimhoff, S. D. *J. Chem. Phys.* **1998**, *109*, 1687–1697.

(32) Mo, Y.; Zhang, Y.; Gao, J. *J. Am. Chem. Soc.* **1999**, *121*, 5737–5742.

(33) Mo, Y.; Gao, J.; Peyerimhoff, S. D. *J. Chem. Phys.* **2000**, *112*, 5530–5538.

(34) Gianinetti, E.; Raimondi, M.; Tornaghi, E. *Int. J. Quantum Chem.* **1996**, *60*, 157–166.

(35) Gianinetti, E.; Vandoni, I.; Famulari, A.; Raimondi, M. *Adv. Quantum Chem.* **1998**, *31*, 251–266.

(36) Raimondi, M.; Famulari, A.; Specchio, R.; Sironi, M.; Moroni, F.; Gianinetti, E. *THEOCHEM* **2001**, *573*, 25–42.

(37) Khaliullin, R. Z.; Head-Gordon, M.; Bell, A. T. *J. Chem. Phys.* **2006**, *124*, 204105.

(38) Khaliullin, R. Z.; Cobar, E. A.; Lochan, R. C.; Bell, A. T.; Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 8753–8765.

(39) Cooper, D. L.; Gerratt, J.; Raimondi, M. *Adv. Chem. Phys.* **1987**, *69*, 319–397.

(40) Wu, W.; Song, L.; Cao, Z.; Zhang, Q.; Shaik, S. *J. Phys. Chem. A* **2002**, *106*, 2721–2726.

(41) Song, L.; Mo, Y.; Zhang, Q.; Wu, W. *J. Comput. Chem.* **2005**, *26*, 514–521.

(42) Hiberty, P. C. *THEOCHEM* **1997**, *398*, 35–43.

(43) Van Lenthe, J. H.; Verbeek, J.; Pulay, P. *Mol. Phys.* **1991**, *73*, 1159–1170.

(44) van Lenthe, J. H.; Dijkstra, F.; Havenith, R. W. A. *Theory Comput. Chem.* **2002**, *10*, 79–116.

(45) Mo, Y.; Song, L.; Lin, Y. *J. Phys. Chem. A* **2007**, *111*, 8291–8301.

(46) Dederichs, P. H.; Bluegel, S.; Zeller, R.; Akai, H. *Phys. Rev. Lett.* **1984**, *53*, 2512–2515.

(47) Behler, J.; Delley, B.; Reuter, K.; Scheffler, M. *Phys. Rev. B* **2007**, *75*, 115409.

(48) Behler, J.; Reuter, K.; Scheffler, M. *Phys. Rev. B* **2008**, *77*, 115421.

(49) Wu, Q.; Cheng, C.-L.; Van Voorhis, T. *J. Chem. Phys.* **2007**, *127*, 164119.

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

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

(52) Gao, J.; Mo, Y. *Prog. Theory Chem. Phys.* **2000**, *5*, 247–268.

(53) Stoll, H.; Wagenblast, G.; Preuss, H. *Theor. Chim. Acta* **1980**, *57*, 169–178.

(54) Nagata, T.; Takahashi, O.; Saito, K.; Iwata, S. *J. Chem. Phys.* **2001**, *115*, 3553–3560.

(55) King, H. F.; Staton, R. E.; Kim, H.; Wyatt, R. E.; Parr, R. G. *J. Chem. Phys.* **1967**, *47*, 1936.

(56) Wu, Q.; Kaduk, B.; Van Voorhis, T. *J. Chem. Phys.* **2009**, *130*, 034109.

(57) Mo, Y. *J. Phys. Chem. A* **2009**, *113*, 5163–5169.

(58) Wu, J. I.; Wannere, C. S.; Mo, Y.; Schleyer, P. v. R.; Bunz, U. H. F. *J. Org. Chem.* **2009**, *74*, 4343–4349.

(59) Mo, Y.; Gao, J. *J. Phys. Chem. A* **2001**, *105*, 6530–6536.

(60) Mo, Y.; Subramanian, G.; Gao, J.; Ferguson, D. M. *J. Am. Chem. Soc.* **2002**, *124*, 4832–4837.

(61) Byun, K.; Mo, Y.; Gao, J. *J. Am. Chem. Soc.* **2001**, *123*, 3974–3979.

(62) Brauer, C. S.; Craddock, M. B.; Kilian, J.; Grumstrup, E. M.; Orilall, M. C.; Mo, Y.; Gao, J.; Leopold, K. R. *J. Phys. Chem. A* **2006**, *110*, 10025–10034.

(63) Mo, Y.; Gao, J. *J. Phys. Chem. B* **2006**, *110*, 2976–2980.

(64) Mo, Y.; Schleyer, P. v. R.; Wu, W.; Lin, M.; Zhang, Q.; Gao, J. *J. Phys. Chem. A* **2003**, *107*, 10011–10018.

(65) Mo, Y.; Wu, W.; Song, L.; Lin, M.; Zhang, Q.; Gao, J. *Angew. Chem., Int. Ed.* **2004**, *43*, 1986–1990.

(66) Mo, Y.; Gao, J. *Acc. Chem. Res.* **2007**, *40*, 113–119.

(67) Marcus, R. A. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1111–1121.

(68) 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.; et al. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

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

(70) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Chem.* **1980**, *58*, 1200.

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

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

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

2716 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Cembran et al.

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

(75) Perdew, J. P. *Phys. Rev. B* **1981**, *23*, 5048.

(76) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 569–582.

(77) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 184112.

(78) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(79) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

(80) Cremer, D.; Filatov, M.; Polo, V.; Kraka, E.; Shaik, S. *Int. J. Mol. Sci.* **2002**, *3*, 604–638.

JCTC Journal of Chemical Theory and Computation

Fully Relativistic, Comparative Investigation of Gold and Platinum Alkyne Complexes of Relevance for the Catalysis of Nucleophilic Additions to Alkynes

Markus Pernpointner*,[†] and A. Stephen K. Hashmi[‡]

Theoretische Chemie, Universität Heidelberg, Im Neuenheimer Feld 229, D-69120 Heidelberg, Germany, and Organisch-Chemisches Institut, Universität Heidelberg, Im Neuenheimer Feld 270, D-69120 Heidelberg, Germany

Abstract: For a range of additions to alkynes gold is known to exhibit a much higher catalytic activity than a corresponding platinum compound. In order to approach the origin of this behavior we first investigate the propyne activation by the gold and platinum catalysts $AuCl_3$ and $PtCl_2(H_2O)$ where both metals possess a d^8 electron configuration and where the catalysts exhibit similar steric effects. Propyne serves as a representative for alkynes. Fully relativistic *ab initio* calculations of these alkyne-catalyst complexes are presented at the Dirac-Hartree−Fock self-consistent field (DHF-SCF), density functional theory (DFT/B3LYP), and Green's function (GF) level in order to properly account for the large relativistic effects of gold and platinum. For the alkyne/catalyst complexes both the perpendicular and in-plane conformations were studied as these possess very similar ground state energies and may easily transform into each other. Strongly varying orbital populations together with sizable energetic and structural differences of the frontier orbitals are found and can be considered as a major source of the differing catalytic activity. These mainly comprise vanishing LUMO densities at the carbon centers in the platinum complex together with increased LUMO energies making a nucleophilic attack harder than in the gold compound. As Green's function calculations show, DFT/B3LYP seems to overestimate correlation contributions leading to an unphysical energetic lowering of many unoccupied orbitals.

Introduction

Unactivated alkynes are not very reactive toward weak nucleophiles. This behavior changes significantly as soon as the C−C triple bond is activated by complexation with a suitable metal catalyst. In this context Hutchings first realized the ability of gold catalysts for the electrophilic alkyne activation[1] making the reaction with weak nucleophiles such as water, alcohols, or amines (Scheme 1) possible. In the meanwhile homogeneous gold catalysis grew into a major field of experimental as well as theoretical research[2−11] opening new pathways for a rich spectrum of organic compounds.

* Corresponding author e-mail: Markus.Pernpointner@ pci.uni-heidelberg.de.
† Theoretische Chemie, Universität Heidelberg.
‡ Organisch-Chemisches Institut.

Scheme 1. Nucleophilic Addition on an Alkyne Activated by a Metal Catalyst
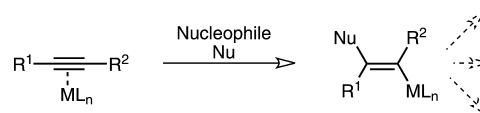
A striking experimental observation now is the considerable activity difference of gold and platinum catalysts as for example described for the gold-catalyzed phenol synthesis in refs 12 and 13. For this reaction calculations at the DFT level have been conducted,[14] and other computational investigations show similar results.[15]

In this work we first focus on the trisubstituted metal catalysts $AuCl_3$ and $PtCl_2(H_2O)$ because of their similarity with respect to *d*-orbital occupation, charge, and steric requirements. They both form a neutral complex with

Scheme 2. Structures (I) and (II) Show the π-like Perpendicular Orientation of the Propyne/Catalyst Complex whereas in the in-Plane Conformations (III) and (IV) a σ-type Bond between the Terminal Carbon Atom and the Metal Is Prevailing

Table 1. Total MP2/RECP Energies (in au) of the Complexes I–IV[a]

$AuCl_3(Pr)_\perp$	$AuCl_3(Pr)_=$	Δ(eV)	$PtCl_2(H_2O)(Pr)_\perp$	$PtCl_2(H_2O)(Pr)_=$	Δ(eV)
−1630.864	−1630.866	0.06	−1231.354	−1231.332	−0.6

[a] The subscripts $\perp$ (=) correspond to the perpendicular (in-plane) conformation in Scheme 2. A positive sign hereby indicates a stability of the planar configuration versus the perpendicular one. (Pr) = propyne.

Table 2. Most Relevant Geometric Parameters for the Complexes I−IV Obtained at the MP2/RECP Level[a]

	$d(MC_1)$	$d(MC_2)$	$d(C_1C_2)$	<($C_1C_2C_3$)	<(HC_1C_2)
Structure I(π)	2.162	2.197	1.242	170.6	166.8
Structure II(π)	2.024	2.040	1.268	161.5	158.0

	$d(MC_1)$	$d(MCl_1)$	$d(C_1C_2)$	$d(C_2Cl)$	<($C_1C_2C_3$)	<(HC_1C_2)
Structure III(σ)	1.972	2.402	1.325	1.825	137.5	126.5
Structure IV(σ)	1.910	2.330	1.329	1.829	140.1	123.7

[a] All bond distances are in Å, all angles in degrees. The symbol (π) stands for the perpendicular orientations where the alkyne π orbitals are engaged in the bond to the metal and (σ) symbolizes the in-plane configuration.

propyne and may be assumed to exhibit very similar properties. As we will show by fully relativistic *ab initio* methods this is, however, not the case. Due to the very small energetic differences both the perpendicular and in-plane orientation of the $AuCl_3/PtCl_2(H_2O)$ and propyne moieties depicted in Scheme 2 had to be considered. At a later stage the positively charged L-Au(I)/propyne complexes will be theoretically investigated taking into account their strongly varying properties of a positive charge and reduced steric requirements.

Both metals exhibit large relativistic effects (see e.g. ref 16 for a comprehensive review over the whole field) which are not only scalar in nature but also comprise considerable spin−orbit coupling. Additionally, electron correlation also contributes to the overall electron density distribution and cannot be treated separately from relativistic effects in heavy systems. For a consistent inclusion of all effects we therefore set out from the Dirac-Coulomb Hamiltonian and calculate the electronic and orbital structure at the Hartree−Fock self-consistent field (HF), density functional (DFT), and one-particle propagator (Green's function) methods. The calculations of the four structures in Scheme 2 were done at their numerically optimized equilibrium geometries.

Geometries and Computational Details

Since a geometry optimization at a four-component correlated level was beyond our computational capabilities, we generated equilibrium geometries for the structures in Scheme 2 at the one-component level. Hereby second-order Møller−Plesset (MP2) theory in combination with relativistic effective core potentials (RECP)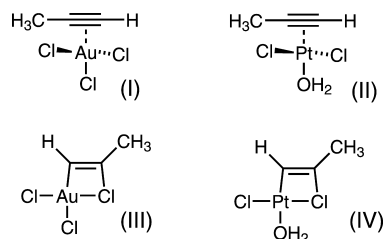[17] from the Stuttgart group was used in order to account for electron correlation and relativistic effects. It should be noted that the effect of spin−orbit coupling on the optimization of equilibrium geometries is neglected at this level of description but due to the closed-shell character of the complexes SO effects on equilibrium structures are assumed to be small. The basis sets were of cc-pVTZ quality for all atoms. It turned out that two energetically similar minimum geometries of the gold and platinum-alkyne complexes exist, namely the perpendicular and the in-plane conformation (see Scheme 2).

Table 1 shows the total energies (in atomic units) for the structures I-IV obtained at the MP2/RECP level of theory. Additionally, the energetic differences (in eV) between the perpendicular and in-plane conformations are given. It turns

out that the energies of the two gold conformers are so similar that both structures may play a role in a reaction path of a nucleophilic attack, or in other words, that there is enough structural flexibility of the gold complex to switch from one orientation to the other during the nucleophilic attack. The two conformers of the platinum complex are energetically more separated with the consequence that a higher activation barrier has to be overcome if one conformation would be more suitable for a nucleophilic attack than the other. For the theoretical analysis of the nucleophilic attack we therefore investigate both geometries with respect to energetic and structural properties. All geometry optimizations for the perpendicular and in-plane conformations converged to C_s symmetry.

In Table 2 we summarize the most important structural parameters of the alkyne/catalyst complexes obtained by the RECP/MP2 geometry optimization. This comprises the metal−carbon bond distances, the alkynylic carbon−carbon distance, and the bending of the carbon backbone.

For further analysis relativistic all-electron calculations at these optimized geometries were performed by using the DIRAC08 program package.[18] Suitable heavy-element basis sets which are adapted to the relativistic change in the electronic structure have to be applied and became available mainly by the work of Dyall.[19−26] From this repository a $(22s19p12d10f1g)$ primitive set for gold and platinum was chosen. The basis sets both include one diffuse f and one core-correlating g function. Due to the large system size the recently developed infinite-order two-component (IOTC)[27] method was employed yielding nearly identical properties for the valence orbitals as a genuine four-component treatment. The two-electron spin−orbit contributions were hereby taken into account by the atomic mean-field approximation[28,29] also available in DIRAC08. As a consequence the orbitals

Gold and Platinum Alkyne Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2719**

Table 3. HF and DFT Orbital Populations in Percent of the Perpendicular (⊥) Orientation of the Catalyst/Propyne Moiety Determined for the IOTC and Douglas-Kroll (DKINF) Hamiltonians[a]

method	orbital	$AuCl_3...Pr_\perp$	$PtCl_2(H_2O)...Pr_\perp$
IOTC	HOMO	94.5 Cl p, 3.9 C p	27.4 M d, 44.6 Cl p, 22.6 C p, 1.2 H s
HF	HOMO-1	8.6 M d, 88.8 Cl p	20.1 M d, 43.8 Cl p, 28.8 C p, 1.7 H s
DKINF	HOMO	95.2 Cl p, 3.8 C p	22.2 M d, 36.3 Cl p, 35.8 C p, 1.9 H s
HF	HOMO-1	8.3 M d, 88.7 Cl p, 2.2 C p	26.1 M d, 54.7 Cl p, 15.5 C p, 1.0 H s
IOTC	HOMO	94.7 Cl p, 4 C p	42.3 M d, 47.5 Cl p, 7.3 C p
DFT	HOMO-1	14.5 M d, 82.4 Cl p	34.1 M d, 42.2 Cl p, 18.1 C p, 1.0 H s
DKINF	HOMO	95.1 Cl p, 2.4 C p	37.7 M d, 59.6 Cl p, 1.1 C p
DFT	HOMO-1	12.3 M d, 86.1 Cl p	39.4 M d, 33.2 Cl p, 23.8 C p

[a] Contributions below one percent are not listed; M stands for metal which is either Au or Pt with the corresponding Mulliken charges of 0.815 and 0.731, respectively.

(spinors) become relaxed with respect to spin−orbit coupling also in the absence of the small-component basis which greatly reduces the numerical effort of the calculation.

In order to distinguish spin−orbit coupling effects from scalar relativistic effects the infinite-order Douglas-Kroll Hamiltonian (DKINF) was employed including relativity only at the scalar level. This is of great importance for the validation of one-component methods which consider only scalar relativistic effects by applying a RECP. Extensive potential energy surface and reaction path calculations can normally not be done at the four-component or even at the IOTC level.

The HOMO and LUMO orbitals obtained by IOTC and DKINF Hamiltonians were then visualized at the HF and DFT level (see Fossgaard et al. for a four-component realization of DFT[30]) in order to elucidate differences in the electronic structure of the gold and platinum complexes. For all relativistic calculations the B3LYP functional[31] was employed. Since DFT often uses parametrized functionals which may lead to constant shifts of the occupied Kohn−Sham orbital energies (see below) the fully relativistic propagator method[32,33] was applied yielding parameter-free correlation corrections to the occupied orbitals in the outer valence space. Propagator-based methods give access to final state distributions resulting from electron detachment/attachment or excitation processes (see for example refs 34−38 for an overview). This can be applied to determine the energies of the occupied valence orbitals including electron correlation and relativistic effects by calculating the corresponding ionization spectra of the complexes. As long as the one-particle picture remains valid (which is in general the case for valence ionizations) the final state energies reflect the positions of the occupied orbitals very accurately. If no electron correlation were taken into account the description would yield the Koopmans energies, but the propagator introduces the sought correlation corrections. It should be mentioned that corrections to the Koopmans energies can be obtained for occupied orbitals only. In this respect the propagator method is very useful to compare *ab initio* correlation shifts to DFT shifts (see discussion below). The numerical algorithm necessary for the solution of the propagator is based on the algebraic diagrammatic construction (ADC).[39]

Results and Discussion

The C≡C triple bond in the complexes is elongated from 1.214 Å in free propyne (optimized geometry at the MP2

level) to 1.242 Å (I), 1.268 Å (II), 1.325 (III), and 1.329 in (IV) clearly indicating an activation of the triple bond in the π-type complexes (I, II) and a substantial change to C−C double-bond (sp^2) character in the σ homologues (III, IV). The transition from a sp to a sp^2 hybridization on carbon is also reflected by a pronounced reduction of the $C_1C_2C_3$ bond angle from 170.6/161.5 degrees (π) to 137.5/140.1 degrees (σ). The HC_1C_2 angles in the σ complexes are even smaller and reflect the sp^2 character of the connected carbon atom. Despite the significant differences in the carbon hybridization and bonding type to the metal atom (σ/π) the total energies are unexpectedly similar as can be seen from the values given in Table 1.

The relativistic stabilization of the $6s$ shell together with a relativistic destabilization of the $5d$ shell in gold and platinum leads to a decreased $5d/6s$ gap enabling larger participation of the $5d$ orbitals in the chemical bonding. In gold, for example this energy gap amounts to 1.150 eV compared to the nonrelativistic value of 5.301 eV.[40] We therefore first investigate the metal $5d$ orbital contributions to the HOMO and HOMO-1 by a population analysis[41] in both complexes. It is known that the absolute values obtained by a Mulliken analysis exhibit some basis set dependency (see for example refs 42 and 43 for a good discussion of this problem), but the procedure is appropriate for comparative purposes as long as basis sets of the same quality and structure are employed for all species under investigation. Afterwards, the effect of electron correlation and spin−orbit splitting on orbital energies and densities will be analyzed.

Metal Populations in the Occupied Frontier Orbitals

We first concentrate on the perpendicularly oriented π-type complexes (I) and (II) of Scheme 2. From Table 3 it can be seen that despite their neighborhood in the periodic table the gold and platinum d orbital participations in the HOMOs of the complexes are substantially different at the HF and DFT level for both the IOTC and DKINF Hamiltonians.

In the transition metals Sc - Zn the $3d$ orbitals are more compact than in the corresponding $4d$ and $5d$ series and show much less participation in the outer valence space than the higher homologues.[44] Despite this, there is no Au $5d$ contribution to the HOMO even at the DFT level including electron correlation and only dominating chlorine p and weak carbon p character is found for the HOMO. The gold $5d$

Table 4. HF and DFT Orbital Populations in Percent of the in-Plane (=) Orientation of the Catalyst/Propyne Moiety Determined for the IOTC and Douglas-Kroll (DKINF) Hamiltonians[a]

method	orbital	AuCl₃...Pr=	PtCl₂(H₂O)...Pr=
IOTC	HOMO	7.6 M d, 65.3 Cl p, 21.9 C p	25.3 M d, 8.3 Cl p, 53.2 C p, 3.6 H s
HF	HOMO-1	2.5 M d, 91.8 Cl p, 2.3 C p	6 M s, 50.2 M d, 38.5 Cl p, 1.3 C p
DKINF	HOMO	7.6 M d, 66.3 Cl p, 22.2 C p	21.2 M d, 8.1 Cl p, 59.1 C p, 3.9 H s
HF	HOMO-1	2.9 M d, 95 Cl p	39.9 M d, 58.5 Cl p
IOTC	HOMO	11.1 M d, 77.8 Cl p, 6.7 C p	59.7 M d, 20.5 Cl p, 11.4 C p
DFT	HOMO-1	5.7 M d, 89.1 Cl p, 1.6 C p	7.3 M s, 56.9 M d, 9.8 Cl p, 18.5 C p, 1.6 H s
DKINF	HOMO	9.7 M d, 80.2 Cl p, 7.9 C p	50.9 M d, 29.8 Cl p, 15.2 C p
DFT	HOMO-1	6.4 M d, 91.6 Cl p	48.9 M d, 6.7 Cl p, 28.7 C p, 1.9 H s

[a] Contributions below one percent are not listed; M stands for metal which is either Au or Pt with the corresponding Mulliken charges of 0.836 and 0.622, respectively.

contributions just appear moderately in the HOMO-1. This is in stark contrast to the situation of the platinum complex where the metal and carbon contributions together make up approximately 50% of the HOMO and HOMO-1 populations. This pronounced structural difference in the outer valence orbitals will also lead to modified LUMO densities giving rise to different affinities to the nucleophile (see next section for the density plots and the discussion of the energetic factors).

The next issue to be addressed is the influence of spin−orbit coupling on the populations. As mentioned before the IOTC Hamiltonian includes spin−orbit coupling for the valence orbitals to the same accuracy as a fully four-component treatment,[27] whereas the infinite order Douglas-Kroll ansatz (DKINF) only treats scalar relativistic effects. One can therefore extract information about the influence of SO-coupling by comparing results obtained by these two Hamiltonians. In the gold complex hardly any effect of SO-coupling is observed at the HF level of theory, and inclusion of electron correlation via DFT does not introduce much of an alteration. This result points to a low sensitivity of the Au-propyne complex toward SO-coupling justifying a scalar relativistic treatment. In the platinum complex, however, we observe significantly larger population changes when going from the IOTC to the DKINF Hamiltonian at both levels of electronic structure treatment (HF, DFT). As a consequence, a purely scalar relativistic description for the platinum complex is not accurate enough for the correct reproduction of the orbital populations but may still be useful if orbital energies and HOMO/LUMO densities change only negligibly upon inclusion of SO coupling.

Due to the sensitivity of transition metal d electrons toward electron correlation one may expect substantial changes in the corresponding d populations as soon as a DFT treatment is applied for the electronic structure. This is indeed the case for the platinum-propyne HOMO and HOMO-1 using the IOTC and DKINF description of relativity. For both Hamiltonians a substantial increase of the Pt 5d population is observed during the transition from a HF to a DFT treatment (for example, the largest change of 15.5% occurs for the HOMO obtained by the DKINF Hamiltonian). For the gold compound, however, neither the HOMOs acquire d character (in fact, the overall populations hardly change at all) nor is the d increase in the HOMO-1 as pronounced as in the platinum case (a maximum change of 5.9% occurs in the HOMO-1/IOTC case). A significantly lower participation of the Au 5d orbitals in the valence space and a reduced

sensitivity of these orbitals toward electron correlation may reflect a more compact d orbital structure in the gold complex. Another observation is the substantial depletion of carbon character in the platinum complex HOMOs in favor of metal d contributions upon inclusion of electron correlation via DFT.

For the perpendicularly oriented complexes (I) and (II) we summarize our findings as follows: The populations of the gold complex show only a weak sensitivity to spin−orbit coupling and electron correlation, and the d orbitals play a less important role for the highest and second highest occupied molecular orbital in all treatments. In contrast, the carbon and platinum d populations change significantly upon inclusion of electron correlation and SO coupling leading to altered LUMO structures as well. The pronounced platinum 5d participation in the outer valence orbitals was also observed in a number of previously calculated dianionic complexes of the PtX₄,₆²⁻ type (X=F, Cl) where one would normally expect a prevailing ligand population.[45]

We next analyze the orbital populations of the in-plane structures (III) and (IV) of Scheme 2 that can be found in Table 4.

At first we now observe a gold 5d HOMO contribution between 7.6% and 11.1% for all Hamiltonians and methods which is in contrast to the results for the perpendicularly oriented Au/Pr complex. Despite the altered overall populations the same insensitivity to SO coupling is visible in the gold complex comparing the results for the IOTC and DKINF Hamiltonian. However, switching from HF to DFT leads to a remarkable reduction of the HOMO carbon contribution and only to a minor increase of the gold 5d population. As it was the case for the perpendicular complexes the platinum species also exhibits a higher sensitivity to SO coupling and a very pronounced change of the 5d character including electron correlation both in the HOMO and HOMO-1. From the results above we observe substantial differences between the gold and platinum complexes both in the perpendicular as well as the in-plane configuration which will have an impact on the affinity toward a nucleophile. Structural factors are mainly based on the different composition of the outer valence orbitals.

The AuCl₃ and PtCl₂(H₂O) fragments of the compounds (I) and (II) have also been calculated separately in order to reveal similar trends in the isolated species as occurring in the complexes. Considering the metal d populations we observe the same pronounced Pt d orbital participation to the three highest occupied orbitals, namely 26.7% (HOMO),

Gold and Platinum Alkyne Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2721**

Table 5. Orbital Energies at Various Levels of Theory in eV of (a) the Perpendicular Au/Pr Complex and (b) the Perpendicular Pt/Pr Complex

orbital	IOTC/ HF	DKINF/ HF	IOTC/ DFT	DKINF/ DFT	IOTC/ ADC(3)
(a) AuCl$_3$...Pr$_\perp$					
LUMO+3	1.442	1.452	−0.583	0.582	-
LUMO+2	1.045	1.046	−1.119	−1.120	-
LUMO+1	0.765	0.766	−2.014	−2.001	-
LUMO	−0.220	−0.288	−3.820	−3.845	-
HOMO	−11.072	−11.075	−7.627	−7.630	−10.218
HOMO−1	−11.456	−11.524	−7.796	−7.850	−10.646
GAP	10.852	10.786	3.807	3.785	-
(b) PtCl$_2$(H$_2$O)...Pr$_\perp$					
LUMO+3	2.151	2.152	−0.105	−0.112	-
LUMO+2	1.682	1.690	−0.630	−0.629	-
LUMO+1	1.445	1.445	−1.008	−1.014	-
LUMO	1.173	1.174	−1.800	−1.807	-
HOMO	−9.844	−9.959	−6.275	−6.427	−8.676
HOMO−1	−10.119	−10.074	−6.638	−6.506	−9.060
GAP	11.018	11.133	4.475	4.620	-

14.3% (HOMO-1), and 18.9% (HOMO-2). In AuCl$_3$ the corresponding *d* character is significantly lower and amounts to 4.4% (HOMO), 7.7% (HOMO-1), and 5.2% (HOMO-2), a behavior which is resembled in the complexes. If one compares the Mulliken charges of the metal centers in the free catalyst and in the complex a small depletion of electron density is observable (0.81 → 0.95 for the Au center and 0.73 → 0.82 for the Pt center). The focus in this work, however, predominantly lies on the analysis of the corresponding alkyne/catalyst complexes and their reactivity toward a nucleophile. For an in-depth analysis of the isolated species we therefore refer to e.g. ref 46.

Energetics and HOMO/LUMO Structures

Considering the orbital energies and structures we first focus on the perpendicular complexes. Table 5a,b contains the corresponding orbital energies for all Hamiltonians and methods. At first we look at the influence of SO coupling on the orbital energies by forming a total average deviation $\Delta\varepsilon_{av}$ spanning the range HOMO-1... LUMO+3. $\Delta\varepsilon_{av}$ is calculated as the sum of all differences $|\varepsilon_{IOTC}^{HF} - \varepsilon_{DKINF}^{HF}|$ and $|\varepsilon_{IOTC}^{DFT} - \varepsilon_{DKINF}^{DFT}|$ divided by the number of orbitals taken into account. For the gold complex we obtain a quite low average deviation of $\Delta\varepsilon_{av} = 0.021$ eV which is not too surprising due to the small population changes upon SO coupling. Despite the considerably larger alterations in the orbital populations $\Delta\varepsilon_{av}$ for the platinum complex increases only up to 0.04 eV where the largest contribution stems from the HOMOs. The direct influence of SO coupling on the highest occupied and lowest virtual orbitals is therefore small.

An important fact is the negative first LUMO energy in the Au/Pr complex at the IOTC/HF and DKINF/HF level reflecting a high affinity toward a nucleophile and the capability to stabilize an excess electron whereas the Pt/Pr complex does not possess a negative LUMO energy. This situation changes pronouncedly when electron correlation is taken into account via DFT. At first the HOMO/LUMO gap is substantially reduced from 10.9 to 3.8 eV in the gold complex and from 11.0 to 4.5 eV in the platinum complex

stressing large electron correlation effects. An even more striking fact, however, is now the decrease of all LUMO energies below zero which would indicate a substantially increased affinity of both complexes toward nucleophiles and a possibility to bind even more than one excess electron. According to the lower-lying LUMOs of the Au/Pr complex this species still forms an even stronger bond to the nuclophile than the platinum complex, but four negative LUMO energies for a neutral closed shell system do not seem to be very reliable. This behavior most probably has its origin in the difficult interpretation of DFT Kohn−Sham (KS) orbital energies.

The Kohn−Sham (KS) formulation of DFT[47] produces one-particle functions (orbitals or spinors in the relativistic DFT) that incorporate electron exchange and correlation effects through the corresponding functionals. However, the interpretation of these orbitals is not undisputed. On the one hand they are seen as purely mathematical constructs in order to reproduce the total electronic density ρ[48] where on the other hand a physical meaning is attributed to them due to the corresponding physical nature of the Kohn−Sham potential.[49,50] If one looks at the occupied orbitals, only the KS-HOMO energy has the physical meaning of the (negative) lowest vertical ionization potential,[51−57] whereas the other are seen as merely auxiliary quantities.[48] Building on the empirical observation that approximate Kohn−Sham energies for the occupied orbitals exhibit a substantial but nearly constant shift with respect to the experimental ionization energies,[58,59] Chong, Gritsenko, and Baerends developed the concept of relaxed vertical ionization potentials (VIPs) and showed a very good agreement of those to the KS orbital energies.[60] In addition, Gritsenko, Braïda, and Baerends[61] analyzed the connection between KS and Dyson orbital theories in the ionization process being relevant for a detailed understanding of the $\varepsilon - I$ relations. In order to estimate this constant shift of the KS orbital energies in our systems we performed fully relativistic propagator calculations of the HOMO energies at the IOTC/ADC(3) level and observe a substantial overestimation of correlation contributions in the DFT treatment (2.6 eV for the Au/Pt and 2.4 eV for the Pt/Pr complex compared to the IOTC/DFT values). As a consequence, the LUMO energies also incorporate this correlation error making the strongly negative orbital energies at the IOTC and DKINF level less troublesome. It should also be mentioned that the approximate functionals may suffer from shortcomings due to self-interaction errors in particular associated with the electron-rich *d* shells. Considering the related DFT-LUMO densities one therefore has to interpret them with care.

For further analysis it is more comprehensive to look at pictorial representations of the HOMO−LUMO orbital densities instead of tabulating dominant basis function coefficients. We commence by displaying the HOMO densities of the perpendicular gold (left) and platinum (right) complexes (I) and (II) at the IOTC/HF level in Figure 1.

In both complexes the occupied alkynylic π orbital and the chlorine lone pairs are clearly visible. The HOMOs obtained at the IOTC/DFT, DKINF/HF, and DKINF/DFT level of theory look nearly identical to the IOTC/HF HOMO

Figure 1. HOMO orbital densities for the perpendicular Au/Pr (left) and Pt/Pr (right) complex at the IOTC/HF level. The methyl group is directed away from the observer.

Figure 2. LUMO orbital densities of the Au/Pr (left) and Pt/Pr (right) complex at the IOTC/HF level. The different orbital densities at the alkynylic carbon atoms are clearly visible.

for both complexes and need not be displayed separately. One of the most prominent differences between the perpendicularly oriented Au and Pt complexes is the differing LUMO density obtained at the IOTC/HF level (see Figure 2) which provides a reasoning for the ease of a nucleophilic attack on the Au/Pr carbyne centers. In contrast, a nonexistent Pt/Pr LUMO density at these carbon centers will not lead to a good overlap of the nucleophile frontier orbital and the complex. As a consequence, a nucleophilic attack on one of the Pt/Pr carbyne centers is less promoted. Looking at the Mulliken charges of the propyne carbon atoms (see Table 7) reveals that the C_2 atom always bears the highest positive charge favoring a Markovnikov addition regardless of available LUMO densities.

This different behavior toward the nucleophile is in line with the negative (positive) LUMO energy of the gold (platinum) complex. The effect of electron correlation (treated at the DFT level of theory) will certainly influence the LUMO densities which are visualized in Figure 3. There the IOTC/DFT LUMO densities at the carbyne atoms of both complexes now show a close resemblance and do not provide a clear-cut argument for the observed affinity difference. From the considerations above

Table 6. Orbital Energies at Various Levels of Theory in eV of the (a) in-Plane Au/Pr Complex and (b) the in-Plane Pt/Pr Complex

orbital	IOTC/ HF	DKINF/ HF	IOTC/ DFT	DKINF/ DFT	IOTC/ ADC(3)
(a) AuCl$_3$...Pr$_=$					
LUMO+3	1.445	1.442	−0.604	−0.607	-
LUMO+2	1.443	1.438	−1.305	−1.308	-
LUMO+1	0.884	0.880	−1.457	−1.453	-
LUMO	0.107	0.055	−3.613	−3.640	-
HOMO	−10.720	−10.743	−7.236	−7.277	−9.608
HOMO−1	−10.957	−10.920	−7.417	−7.361	−9.907
GAP	10.827	10.798	3.623	3.637	-
(b) PtCl$_2$(H$_2$O)...Pr$_=$					
LUMO+3	1.990	1.988	−0.304	−0.308	-
LUMO+2	1.947	1.954	−0.383	−0.406	-
LUMO+1	1.863	1.860	−0.543	−0.509	-
LUMO	1.264	1.262	−1.107	−1.119	-
HOMO	−9.252	−9.336	−5.980	−6.182	−8.072
HOMO−1	−9.705	−9.932	−6.077	−6.242	−8.244
GAP	10.516	10.598	4.872	5.063	-

one should keep in mind, however, that the DFT HOMO/ LUMO energies bear intrinsic shifts, and the corresponding
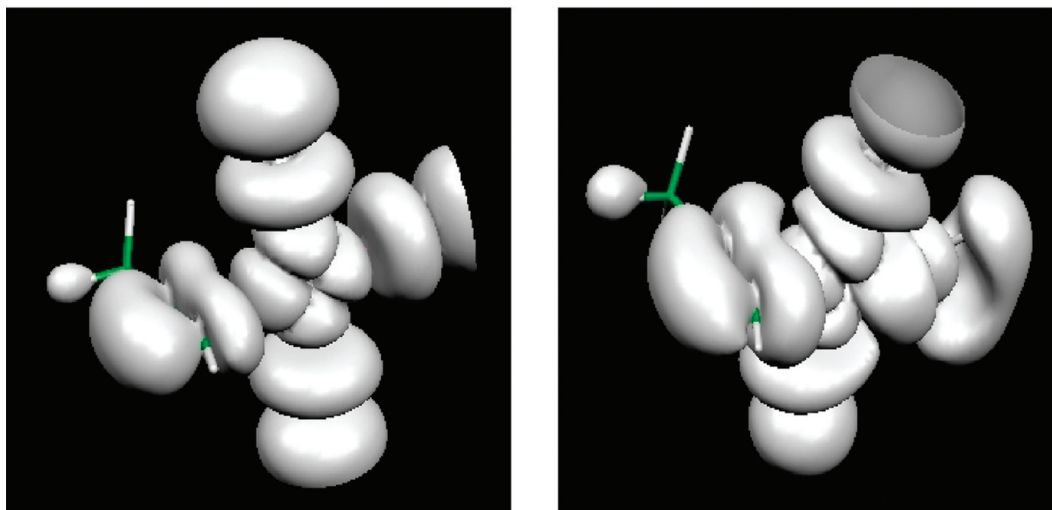
Figure 3. LUMO orbital densities of the Au/Pr (left) and Pt/Pr (right) complex at the IOTC/DFT level. The LUMO density differences are now blurred.

Table 7. Mulliken Charges at the DFT/B3LYP and HF Level Using Relativistic ECPs[a]

method	C_1	C_2	C_3
Au$_\perp$ B3LYP	+0.001	+0.0522	−0.273
HF	−0.049	+0.0529	−0.274
Au$_=$ B3LYP	−0.122	+0.0317	−0.263
HF	−0.155	+0.0319	−0.244
Pt$_\perp$ B3LYP	−0.175	+0.0489	−0.274
HF	−0.221	+0.0393	−0.259
Pt$_=$ B3LYP	−0.226	+0.0201	−0.275
HF	−0.315	+0.0387	−0.253
C_3H_4 HF	−0.199	−0.0747	−0.2752

[a] Au (Pt) stands for the gold (platinum)/propyne complex and $\perp$ (=) for the perpendicular (in-plane) orientation. The charges of free propyne are given in the last line. The metal-induced activation of the C_2 atom toward a nucleophilic attack is clearly visible.

densities may therefore also deviate considerably from the actual situation. For this reason we do not establish our reasoning purely on DFT results.

In the last paragraph of this section we discuss the in-plane complexes whose orbital energies are shown in Table 6a,b at various levels of theory.

Despite the different hybridization of the propyne carbon atoms similar global trends as in the perpendicular configurations can be observed. The HOMO and LUMO energies are again all lower in the Au/Pr complex than in the corresponding Pt/Pr compound, and inclusion of electron correlation leads to a similarly large increase of the HOMO and decrease of the LUMO energies. For the determination of the shift introduced by DFT we again compare the IOTC/DFT HOMO energy in both complexes to the IOTC/ADC(3) value and obtain 2.4 eV for the Au/Pr and 2.1 eV for the Pt/Pr complex. In contrast to the perpendicular gold complex no negative first LUMO energy is obtained at the IOTC/HF and DKINF/HF level. Naturally the question arises if any affinity differences can be associated with the configurational change from perpendicular to in-plane or if it is related to the type of catalyst. To answer this we again visualize the corresponding orbital densities for the in-plane configurations.

In Figure 4 showing the IOTC/HF HOMO densities of the in-plane Au/Pr and Pt/Pr complexes we do not observe a

substantial difference at the alkynylic carbon atoms and very similar plots (not displayed) also result in a scalar relativistic DKINF/HF treatment. From Figure 5 we see the same depletion of LUMO density at the carbyne atoms of the in-plane platinum complex calculated at the IOTC/HF level. This difference is therefore related to the catalyst (gold or platinum center) and not to the orientation of the alkyne moiety. According to the nonexistent HF LUMO density on the carbyne positions the Pt/Pr complex should exhibit a decreased affinity to the nucleophile. The increased affinity of the gold complex is also reflected by a considerably lower LUMO energy of +0.107 eV compared to +1.264 eV in the platinum complex.

The visualizations of the KS HOMOs using the IOTC and DKINF Hamiltonians will not introduce new information and are omitted. In all cases the KS orbital densities at the carbyne centers look very similar to the HF results, and SO coupling does not alter this picture. This is not the case with the LUMO densities depicted in Figure 6 at the IOTC/DFT level going in line with the findings for the perpendicular configurations. As before, the difference in the LUMO densities at the carbon atoms of both complexes is wiped out and does not allow for a clear structural argument with respect to affinities.

From this structural analysis we deduce that the type of catalyst (Au or Pt) influences the electronic structure and the densities at the reactive centers much more than the geometric orientation of the two building blocks even if the latter implies a change of the carbon hybridization. However, due to its negative LUMO energy the perpendicular gold complex clearly alleviates a nucleophilic attack, whereas a reorientation in the platinum compound does not exhibit a comparable behavior. The low energy barrier of 0.06 eV for a transformation between the two orientations of Au/Pr leads to nucleophilic reactions where only the perpendicular type is involved. This may be of relevance for further mechanistic studies of the nucleophilic attack.

In total we can summarize our findings as follows. Spin−orbit coupling has a considerable influence on the orbital populations in the Pt-containing complexes but leaves the outer valence compositions in the corresponding gold species nearly unaltered. Orbital energies of both complexes are hardly modified upon
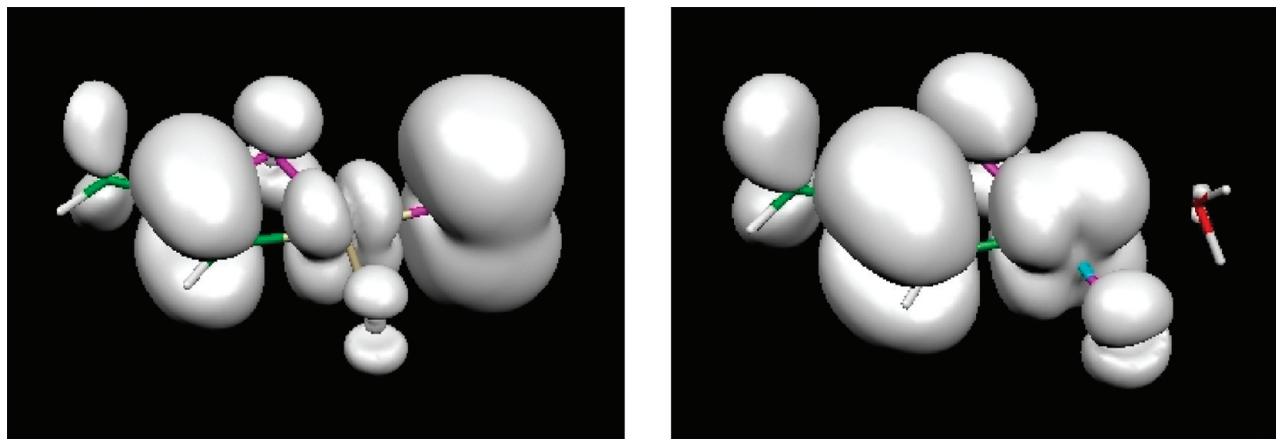
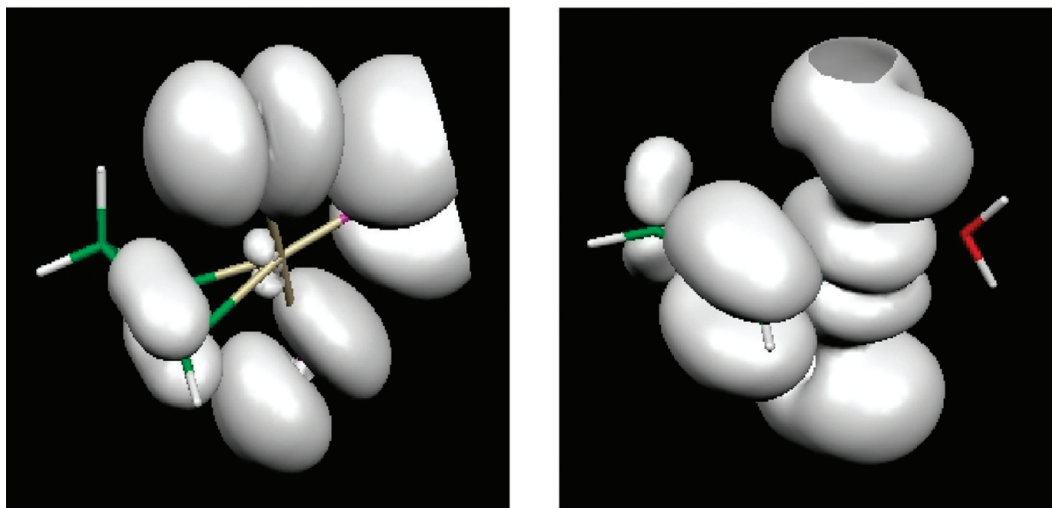**Figure 4.** HOMO orbital densities for the in-plane Au/Pr (left) and Pt/Pr (right) complex at the IOTC/HF level. The methyl group is directed away from the observer.
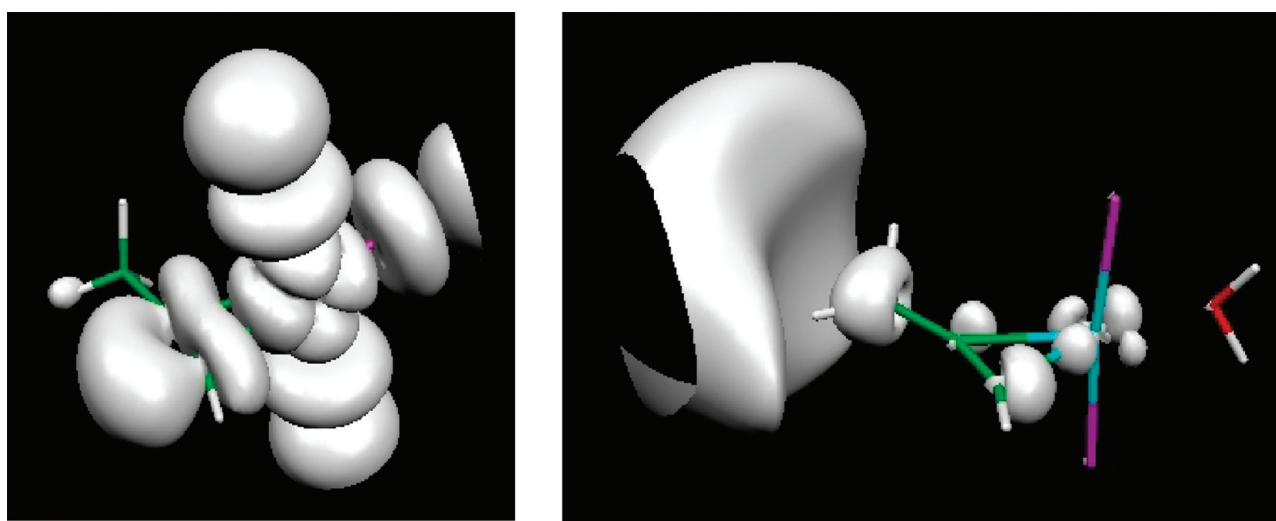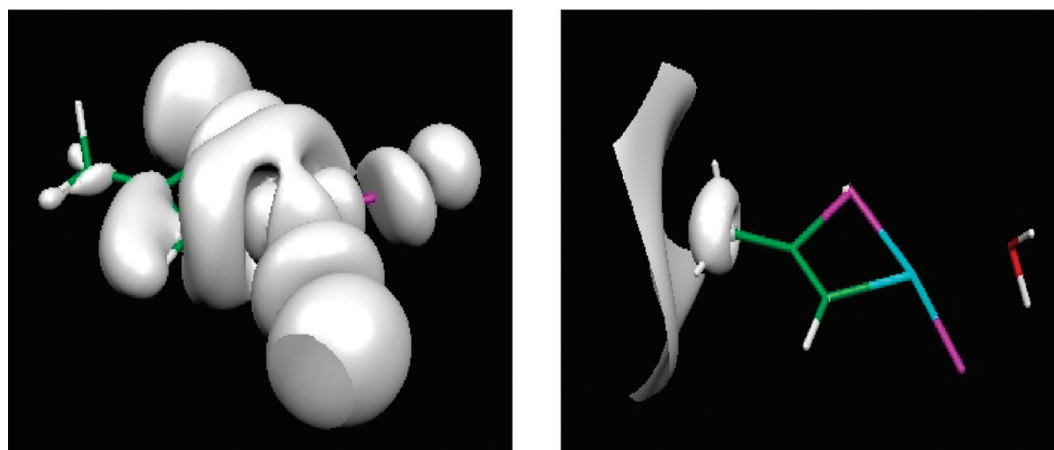

Figure 5. LUMO orbital densities for the in-plane Au/Pr (left) and Pt/Pr (right) complex at the IOTC/HF level.
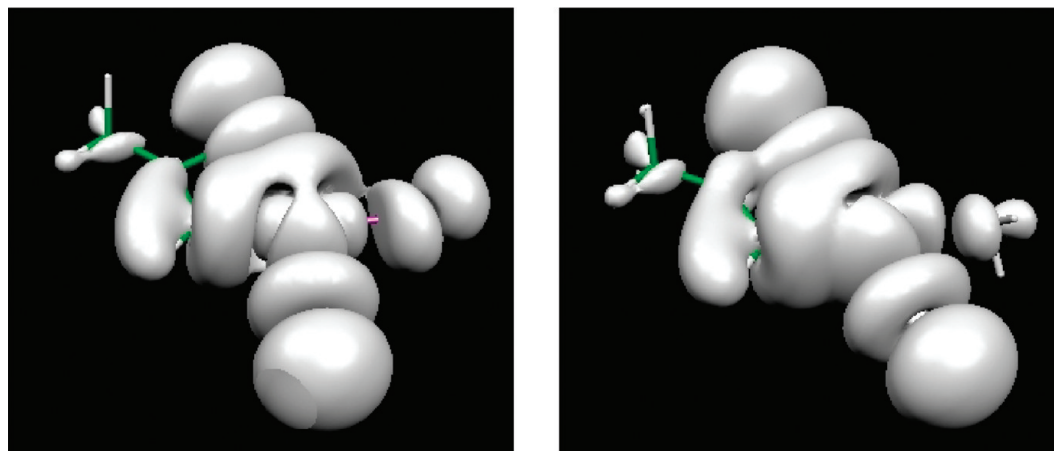
Figure 6. LUMO orbital densities for the in-plane Au/Pr (left) and Pt/Pr (right) complex at the IOTC/DFT level. The methyl group is directed away from the observer.

inclusion of SO coupling which is understandable due to their closed-shell character. Very differing metal *d* populations in the outer valence space lead to large structural alterations in the LUMOs which can be considered as one main ingredient for the different catalytic efficiency. Hereby no carbon LUMO density was found for the platinum complexes making an overlap with a nucleophilic frontier orbital unfavorable. In contrast to that both gold complexes exhibit a high LUMO density together with considerably lower LUMO energies alleviating a nucleophilic attack both from the structural as from

the energetic point of view. The orientation of the propyne moiety (perpendicular or in-plane) is hereby irrelevant. Electron correlation treated at the relativistic DFT level has a major impact on orbital energies, populations, and densities. The interpretation of the DFT results especially with respect to virtual orbital energies is not straightforward as discussed in the text, and relativistic propagator calculations reveal a considerable overestimation of correlation effects by DFT. Therefore the obtained DFT-LUMO densities cannot serve as an unambiguous basis for the structural analysis. For all four species under consideration the

Gold and Platinum Alkyne Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2725**

central carbon atom bears the highest positive charge, and a Markovnikov-type attack is therefore most probable.

References

(1) Hutchings, G. J. *J. Catal.* **1985**, *96*, 292.

(2) Dyker, G. *Angew. Chem.* **2000**, *112*, 4407.

(3) Dyker, G. *Angew. Chem., Int. Ed. Engl.* **2000**, *39*, 4237.

(4) Hashmi, A. S. K. *Gold Bull.* **2003**, *36*, 3.

(5) Hashmi, A. S. K. *Gold Bull.* **2004**, *37*, 51.

(6) Hofmann-Röder, A.; Krause, N. *Org. Biomol. Chem.* **2005**, *3*, 387.

(7) Hashmi, A. S. K. *Angew. Chem.* **2005**, *117*, 7150.

(8) Hashmi, A. S. K. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 6990.

(9) Hashmi, A. S. K.; Hutchings, G. *Angew. Chem.* **2006**, *118*, 8064.

(10) Hashmi, A. S. K.; Hutchings, G. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 7896.

(11) Hashmi, A. S. K. *Chem. Rev.* **2007**, *107*, 3180.

(12) Hashmi, A. S. K.; Weyrauch, J. P.; Rudolph, M.; Kurpejovic, E. *Angew. Chem.* **2004**, *116*, 6707.

(13) Hashmi, A. S. K.; Weyrauch, J. P.; Rudolph, M.; Kurpejovic, E. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 6545.

(14) Martin-Matute, B.; Nevado, C.; Cárdenas, D. J.; Echavarren, A. M. *J. Am. Chem. Soc.* **2003**, *125*, 5757.

(15) Soriano, E.; Marco-Contelles, J. *Organometallics* **2006**, *25*, 4542.

(16) *Relativistic Electronic Structure Theory*; Schwerdtfeger, P., Ed.; Eleesevier B. V.: Amsterdam, 2004.

(17) Figgen, D.; Rauhut, G.; Dolg, M.; Stoll, H. *Chem. Phys.* **2005**, *311*, 227.

(18) Visscher, L.; Jensen, H. J. A.; Saue, T. *Dirac, a relativistic ab initio electronic structure program, Release DIRAC08 (2008)*,with new contributions from Bast, R., Dubillard, S., Dyall, K. G., Ekström, U., Eliav, E., Fleig, T., Gomes, A. S. P., Helgaker, T. U., Henriksson, J., Iliaš, M., Jacob, Ch. R., Knecht, S., Norman, P., Olsen, J., Pernpointner, M., Ruud, K., Sałek, P., Sikkema, J. (see http://dirac.chem.sdu.dk (accessed September 1, 2009)).

(19) Dyall, K. G. *Theor. Chem. Acc.* **1998**, *99*, 366.

(20) Dyall, K. G. *Theor. Chem. Acc.* **2002**, *108*, 335.

(21) Dyall, K. G. *Theor. Chem. Acc.* **2002**, *108*, 365.

(22) Dyall, K. G. *Theor. Chem. Acc.* **2003**, *109*, 284.

(23) Dyall, K. G. *Theor. Chem. Acc.* **2004**, *112*, 403.

(24) Dyall, K. G. *Theor. Chem. Acc.* **2006**, *115*, 441.

(25) Dyall, K. G. *Theor. Chem. Acc.* **2007**, *117*, 483.

(26) Dyall, K. G. *Theor. Chem. Acc.* **2007**, *117*, 491.

(27) Iliaš, M.; Saue, T. *J. Chem. Phys.* **2007**, *126*, 064102.

(28) Hess, B. A.; Marian, C. M.; Wahlgren, U.; Gropen, O. *Chem. Phys. Lett.* **1996**, *251*, 365.

(29) Iliaš, M.; Kellö, V.; Visscher, L.; Schimmelpfennig, B. *J. Chem. Phys.* **2001**, *115*, 9667.

(30) Fossgaard, O.; Gropen, O.; Corral Valero, M.; Saue, T. *J. Chem. Phys.* **2003**, *118*, 10418.

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

(32) Pernpointner, M.; Trofimov, A. B. *J. Chem. Phys.* **2004**, *120*, 4098.

(33) Pernpointner, M. *J. Chem. Phys.* **2004**, *121*, 8782.

(34) Abrikosov, A. A.; Gorkov, L. P.; Dzyaloshinski, I. E. *Methods of Quantum Field Theory in Statistical Physics*; Prentice-Hall: Englewood Cliffs, 1963.

(35) Fetter, A. L.; Walecka, J. D. *Quantum Theory of Many-Particle Systems*; McGraw-Hill: New York, 1971.

(36) Mattuck, R. D. *A Guide to Feynman Diagrams in the Many-Body Problem*; McGraw-Hill: New York, 1967.

(37) Cederbaum, L. S.; Domcke, W. *Adv. Chem. Phys.* **1977**, *36*, 205.

(38) Cederbaum, L. S.; Domcke, W.; Schirmer, J.; von Niessen, W. *Adv. Chem. Phys.* **1986**, *65*, 115.

(39) Schirmer, J.; Cederbaum, L. S.; Walter, O. *Phys. Rev. A* **1983**, *28*, 1237.

(40) Eliav, E.; Kaldor, U.; Ishikawa, Y. *Phys. Rev. A* **1994**, *49*, 1724.

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

(42) Wiberg, K. B.; Rablen, P. R. *J. Comput. Chem.* **1993**, *14*, 1504.

(43) Meister, J.; Schwarz, W. H. E. *J. Phys. Chem.* **1994**, *98*, 8245.

(44) Frenking, G.; Fröhlich, N. *Chem. Rev.* **2000**, *100*, 717.

(45) Pernpointner, M.; Rapps, T.; Cederbaum, L. S. *J. Chem. Phys.* **2008**, *129*, 174302.

(46) Hargittai, M.; Schulz, A.; Réffy, B.; Kolonits, M. *J. Am. Chem. Soc.* **2001**, *123*, 1449.

(47) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

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

(49) Baerends, E. J.; Gritsenko, O. V. *J. Phys. Chem. A* **1997**, *101*, 5383.

(50) Baerends, E. J. *Theor. Chem. Acc.* **2000**, *103*, 1265.

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

(52) Perdew, J. P.; Levy, M. *Phys. Rev. Lett.* **1983**, *51*, 1884.

(53) Levy, M.; Perdew, J. P.; Sahni, V. *Phys. Rev. A* **1984**, *30*, 2745.

(54) Almbladh, C. O.; Pedroza, A. C. *Phys. Rev. A* **1984**, *29*, 2322.

(55) Almbladh, C. O.; von Barth, U. *Phys. Rev. B* **1985**, *31*, 3231.

(56) Perdew, J. P.; Levy, M. *Phys. Rev. B* **1997**, *56*, 16021.

(57) Casida, M. *Phys. Rev. B* **1998**, *59*, 4694.

(58) Baerends, E. J.; Ros, P. *Chem. Phys.* **1973**, *2*, 52.

(59) Politzer, P.; Abu-Awwad, F. *Theor. Chem. Acc.* **1998**, *99*, 83.

(60) Chong, D. P.; Gritsenko, O. V.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 1760.

(61) Gritsenko, O. V.; Braïda, B.; Baerends, E. J. *J. Chem. Phys.* **2003**, *119*, 1937.

JCTC Journal of Chemical Theory and Computation

Benchmark Data for Noncovalent Interactions in HCOOH···Benzene Complexes and Their Use for Validation of Density Functionals

Yan Zhao,* Hou T. Ng, and Eric Hanson

Commercial Print Engine Lab, HP Laboratories, Hewlett-Packard co. 1501 Page Mill Road, Palo Alto, California 94304

Abstract: We present benchmark energetic data for the HCOOH···benzene complexes. The benchmark data were determined by a composite approach based on CCSD(T) calculations. Final binding energies (kcal/mol) are in the range of 1.6−4.8 kcal/mol, and they were used as reference data to test density functionals in the literature. Among the tested local density functionals without empirical dispersion corrections, M06-L is the best performing functional, and M06-L/6-31+G(d,p) gives a mean unsigned error (MUE) of only 0.15 kcal/mol. PBEsol and SOGGA also show promising performance. The best local DFT-D methods are BLYP-D and PBEsol-D, and they give an MUE of 0.15 kcal/mol after removing basis set superposition errors by the counterpoise approach. Empirical dispersion corrections greatly improve the descriptions of noncovalent interactions in HCOOH···benzene dimers. The calculated benchmark data and intermolecular potential are useful for the parametrizations of new force fields and coarse-grained models for chemical species such as the acrylic polymers.

1. Introduction

An understanding of noncovalent interactions between a polar molecule and an aromatic system is of fundamental interest since these interactions are common not only in biological systems[1,2] but also in many areas of chemistry and materials science.[3−7] An important example of this type of interaction, water···benzene dimer, has been extensively studied both experimentally and theoretically.[8−14] Recently Crittenden[15] carried out a systematical and theoretical study of the long-range noncovalent interactions between benzene and a series of hydrides and rare gas atoms. More recently, Imai et al.[16] reported a computational study of the amide-π interactions between formamide (HCONH$_2$) and benzene, and Ottiger et al.[17] performed a theoretical study of the amide-benzene interactions in the 2-pyridone···benzene complexes. In the present study, we focus on understanding the noncovalent interactions between formic acid (HCOOH) and benzene.

The HCOOH···benzene complex can be considered as a model system for understanding the side-chain interactions in acrylic polymers (or resins), because they are commonly being used in the coating and printing industry. Furthermore, the HCOOH···benzene dimer can be viewed as a model system for studying pigment-polymer interactions since common pigments such as those based on Cu-phthalocyanine have benzene rings in their structures, and acrylic polymers have −COOH side-chains. It is important to understand such physical interactions for rational design of future high-performance materials for next-generation printing materials and technology.

The experimental or computational studies of the HCOOH···benzene complex are scarce in the literature. An objective of the present study is to use Møller−Plesset second order perturbation theory (MP2)[18] and coupled cluster theory[19,20] with single and double excitations and a quasiperturbative treatment of connected triple excitations (CCSD(T)) to characterize the HCOOH···benzene complexes and to develop benchmark energetic data for interaction energies in HCOOH···benzene dimers. A further goal of our study is to validate a number of low-cost density functional theory (DFT) methods and to determine if these can describe the energetics of noncovalent interactions in HCOOH···benzene complexes.

* Corresponding author e-mail: yan.zhao3@hp.com.

Noncovalent Interactions in HCOOH···Benzene Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2727**
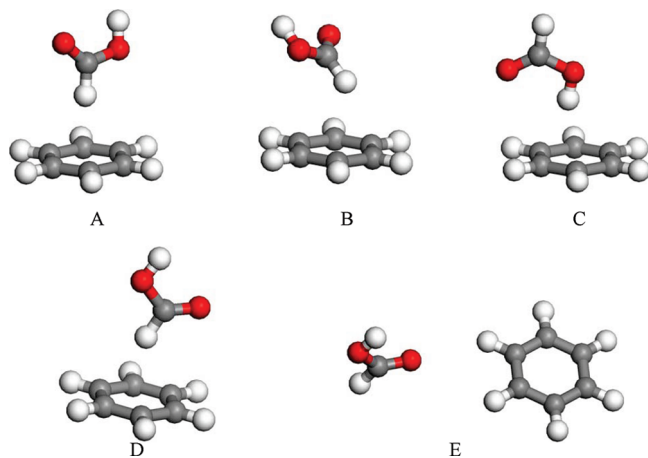
Figure 1. Five stationary points on the potential energy surface of an HCOOH···benzene complex.

2. Computational Methods

We have employed MP2[18] and the 6-311+G(2df,2p) basis set[21] to optimize the geometries of the HCOOH···benzene complexes without any geometrical constraints. We started with many initial guesses of the dimer configurations for optimizations, following the procedure described in a previous paper for HCONH$_2$···benzene.[16] Five stationary points have been located as the minimum structures on the MP2/6-311+G(2df,2p) potential energy surface, and they are shown in Figure 1.

We estimated the CCSD(T) complete basis limit binding energies in these five complexes as follows

$$D_e(\text{CCSD(T)/CBS}) = D_e(\text{MP2/CBS}) + \\ D_e(\text{CCSD(T)/SB}) - D_e(\text{MP2/SB}) \quad (1)$$

where CBS denotes the complete-basis limit, and SB denotes a small basis (aug-cc-pVDZ). The MP2/CBS energies are extrapolated by

$$E^{\text{MP2}}(n) = E_\infty^{\text{MP2}} + An^{-3} \quad (2)$$

where n represents the highest angular momentum in an augmented correlation-consistent basis set, $n = 3$ for the aug-cc-pVTZ basis, and $n = 4$ for the aug-cc-pVQZ basis sets. We also examine a lower-cost MP2/IB extrapolation using the aug-cc-pVDZ and aug-cc-pVTZ basis sets. IB denotes the separate extrapolation[22-24] of Hartree−Fock and MP2 correlation energies to the infinite-basis (IB) limit. The Hartree−Fock energies are extrapolated by

$$E^{\text{HF}}(n) = E_\infty^{\text{HF}} + A^{\text{HF}}n^{-\alpha} \quad (3)$$

and the MP2 correlation energies are extrapolated by

$$E^{\text{cor}}(n) = E_\infty^{\text{cor}} + A^{\text{cor}}n^{-\beta} \quad (4)$$

The parameters α and β in eqs 3 and 4 are determined in a previous paper;[24] the value used for α is 4.93, and that for β is 2.13. Combining MP2/IB with the (CCSD(T)-MP2) correction evaluated at the aug-cc-pVDZ basis set, one can define an estimated CCSD(T)/IB binding energies

$$D_e(\text{CCSD(T)/IB}) = D_e(\text{MP2/IB}) + D_e(\text{CCSD(T)/SB}) - \\ D_e(\text{MP2/SB}) \quad (5)$$

This composite approach defined in eq 5 has been shown[23] to give very good agreement with experiments for the predictions of binding energies in the H$_2$O···benzene and NH$_3$···benzene systems, provided that counterpoise (Cp) corrections for basis set superposition error (BSSE) are included. Therefore we performed all benchmark calculations with Cp corrections[25-27] for BSSE.

We have tested density functionals with the MP2/6-311+G(2df,2p) geometries. The tested functionals can be classified according to various rungs of "Jacob's ladder".[28] The lowest rung is the local spin density approximation (LSDA), in which the density functional depends only on spin densities, and the second rung is the generalized gradient approximation (GGA, in which the density functional depends on spin densities and their reduced gradient). The third-rung functionals, meta-GGAs, also depends on the spin kinetic energy densities. The fourth rung is hyper GGA,[28] which employs full or partial exact Hartree−Fock (HF) exchange. There are two types of hyper GGAs on the fourth rung, namely the hybrid GGAs (HF + GGA) and hybrid meta-GGAs (HF + meta-GGA). In this work, the tested conventional functionals include one LSDA,[29-31] 9 GGAs (BLYP,[32,33] BP86,[32,34] G96LYP,[33,35] HCTH,[36] mPWPW,[37] OLYP,[33,38] PBE,[39] PBEsol,[40] and SOGGA[41]), three meta-GGAs (M06-L,[42] TPSS,[43,44] and VSXC[45]), 12 hybrid GGAs (B3LYP,[46] B97-1,[36] B97-2,[47] B98,[48] BHandH,[49] BHandH-LYP,[49] MPW1K,[50] mPW1PW,[37] MPW3LYP,[33,37,51] O3LYP,[38,52] PBE0,[39] X3LYP[53]), and 9 hybrid meta-GGAs (BMK,[54] PW6B95,[55] PWB6K,[55] M05,[56] M05-2X,[57] M06,[58] M06-2X,[58] M06-HF,[59] and TPSSh[43,44]).

Recently several DFT-D schemes have been developed[60-63] by augmenting the conventional DFT energy with a damped empirical dispersion term ($-C_6R^{-6}$), these DFT-D schemes have been shown to give much improved performance for noncovalent interactions. In the present study, we test B97-D,[60] TPSS-D,[60] BP86-D,[60] B3LYP-D,[60] PBE-D,[60] PBEsol-D,[61] ωB97X-D,[62] PBE0-D,[63] BMK-D,[63] M06-L-D,[63] M06-D,[63] and SOGGA-D. SOGGA-D is defined in the present study by using the same empirical dispersion corrections as PBEsol-D.

In the Grimme's dispersion damping function[60] for DFT-D calculations, the van der Waals radii are derived from the radius of the 0.01 a_0^{-3} electron density contour from ROHF/TZV calculations of the atoms in the ground state, and Grimme used a scale factor of 1.1 to scale these radii. The van der Waals radii in Table 1 of ref 60 are scaled values. In PBEsol-D (also in SOGGA-D), the scale factor is increased to 1.42.[61] We implemented the new scale factor in the locally modified Gaussian09 program for PBEsol-D and SOGGA-D calculations.

All DFT calculations have been performed with the locally modified GAMESS,[64] NWChem,[65] and Gaussian09[66] programs.

3. Results and Discussion

3.1. Stationary Points on the Potential Energy of HCOOH···Benzene Noncovalent Complexes. Figure 1 shows the five minimum structures located by the MP2/6-

Table 1. Benchmark Results (in kcal/mol) for the Binding Energies[a] of the HCOOH···Benzene Complexes

complex	HF/IB[b]	ΔMP2/IB[c]	MP2/IB[d]	MP2/CBS[e]	Δ(CCSD(T)-MP2)[f]	CCSD(T)/IB[g]	CCSD(T)/CBS[h]
A	−0.63	4.40	3.77	3.60	−0.45	3.32	3.15
B	−1.20	5.25	4.05	3.88	−0.58	3.48	3.30
C	0.18	5.51	5.69	5.44	−0.65	5.04	4.79
D	−0.82	4.69	3.87	3.70	−0.47	3.41	3.23
E	0.19	1.54	1.73	1.68	−0.06	1.67	1.62

[a] The binding energy D_e is defined in the present study as the equilibrium dissociation energy of the complexes dissociated into HCOOH and benzene. All calculations employed MP2/6-311+G(2df,2p) geometries. [b] Extrapolated ΔE_{HF} with eq 3 and the aug-cc-pVDZ and aug-cc-pVTZ basis sets. [c] The extrapolated ($\Delta E_{MP2}-\Delta E_{HF}$) results with eq 4 and the aug-cc-pVDZ and aug-cc-pVTZ basis sets. [d] The sum of HF/IB and ΔMP2/IB. [e] The extrapolated ΔE_{MP2} results with eq 2 and the aug-cc-pVTZ and aug-cc-pVQZ basis sets. [f] This term is evaluated with the aug-cc-pVDZ basis set. [g] This column gives D_e estimated using eq 5, i.e. the summation of values in the MP2/IB and Δ(CCSD(T) - MP2) columns. [h] This column gives D_e estimated using eq 1, i.e. the summation of values in the MP2/CBS and Δ(CCSD(T) - MP2) columns

Table 2. Binding Energies and Mean Errors (in kcal/mol) for Local Density Functionals[a]

method	A	B	C	D	E	MSE[b]	MUE[c]
best estimate[d]	3.15	3.30	4.79	3.23	1.62		
M06-L	2.60	2.98	4.60	2.71	1.19	−0.40	0.40
SOGGA	2.47	2.38	4.56	2.53	1.29	−0.57	0.57
PBEsol	2.29	2.20	4.30	2.35	1.18	−0.76	0.76
LSDA	4.52	4.81	7.12	4.76	2.48	1.52	1.52
PBE	1.27	1.02	2.85	1.25	0.91	−1.76	1.76
TPSS	0.71	0.30	2.07	0.61	0.48	−2.39	2.39
mPWPW	0.42	0.01	1.80	0.31	0.32	−2.65	2.65
τ-HCTH	0.07	−0.50	1.15	−0.14	0.45	−3.01	3.01
HCTH	−0.10	−0.63	0.96	−0.27	0.92	−3.04	3.04
BP86	−0.05	−0.45	1.32	−0.13	−0.16	−3.11	3.11
BLYP	−0.73	−1.25	0.29	−0.85	−0.02	−3.73	3.73
OLYP	−2.45	−3.38	−1.72	−2.77	−0.66	−5.41	5.41
G96LYP	−2.87	−3.67	−2.03	−3.08	−1.76	−5.90	5.90
VSXC	9.14	11.65	12.71	10.72	4.21	6.47	6.47

[a] Tested with the 6-311+G(2df,2p) basis set and MP2/6-311+G(2df,2p) geometries. [b] MSE denotes mean signed error (same as mean deviation, MD). [c] MUE denotes mean unsigned error (same as mean absolute deviation, MAD). [d] Taken from the CCSD(T)/CBS results of Table 1.

311+G(2df,2p) level of theory. The first four complexes (A-D) are π-hydrogen bonded complexes, in which the hydrogen bond acceptor is the π-cloud on the benzene molecule. The fifth complex (E) is bound by the interaction between the sp^2 oxygen in HCOOH and two hydrogens of benzene.

3.2. Benchmark Calculations. Table 1 gives the benchmark results for the noncovalent interaction energies in the five HCOOH···benzene complexes.

Table 1 shows that Hartree−Fock (the HF/IB column of Table 1) theory severely underestimates the noncovalent interactions in the HCOOH···benzene complexes because HF theory gives small or repulsive (or negative) binding energies (at the MP2/6-311+G(2df,2p) geometries). The attractive interactions in these complexes are mostly due to medium-range correlation energies, which are absent in Hartree−Fock theory.

As shown in Table 1, the MP2/IB calculations agree with MP2/CBS within 0.2 kcal/mol, but MP2/IB is a much less expensive method than MP2/CBS. For large systems, MP2/IB is an efficient alternative to the MP2/CBS approach.

Table 3. Binding Energies and Mean Errors (in kcal/mol) for Hybrid Density Functionals[a]

method	A	B	C	D	E	MSE[b]	MUE[c]
best estimate[d]	3.15	3.30	4.79	3.23	1.62		
PWB6K	2.89	2.97	4.63	3.09	1.62	−0.18	0.18
M05-2X	3.32	3.59	5.43	3.43	1.61	0.26	0.26
M06	2.54	2.90	4.49	2.63	1.26	−0.45	0.45
M06-2X	3.54	4.04	5.68	3.77	1.59	0.50	0.52
PW6B95	2.17	2.18	3.78	2.35	1.22	−0.88	0.88
M05	1.81	1.79	3.39	1.93	1.56	−1.12	1.12
M06-HF	4.44	4.92	6.60	4.59	1.77	1.24	1.24
PBE0	1.58	1.34	3.20	1.55	0.98	−1.49	1.49
B97-1	1.55	1.34	3.00	1.54	1.15	−1.50	1.50
BHandH	4.66	4.94	7.01	4.89	2.72	1.63	1.63
MPW1K	1.36	1.07	2.89	1.29	0.74	−1.75	1.75
B98	1.20	0.91	2.57	1.14	0.94	−1.87	1.87
BMK	1.09	1.08	2.78	1.31	0.25	−1.92	1.92
MPW3LYP	1.10	0.80	2.42	1.05	1.13	−1.92	1.92
BHandHLYP	1.08	0.77	2.41	1.03	0.98	−1.97	1.97
τ-HCTHh	1.04	0.77	2.51	0.98	0.64	−2.03	2.03
mPW1PW	0.95	0.60	2.42	0.86	0.56	−2.14	2.14
TPSSh	0.87	0.48	2.27	0.78	0.54	−2.23	2.23
TPSS1KCIS	0.81	0.42	2.11	0.72	0.66	−2.28	2.28
X3LYP	0.74	0.40	2.04	0.68	0.83	−2.28	2.28
B97-2	0.27	−0.16	1.53	0.14	0.31	−2.80	2.80
B3LYP	0.20	−0.21	1.42	0.12	0.46	−2.82	2.82
O3LYP	−1.59	−2.38	−0.67	−1.85	−0.27	−4.57	4.57

[a] Tested with the 6-311+G(2df,2p) basis set and MP2/6-311+G(2df,2p) geometries. [b] MSE denotes mean signed error (same as mean deviation). [c] MUE denotes mean unsigned error (same as mean absolute deviation, MAD). [d] Taken from the CCSD(T)/CBS results of Table 1.

Noncovalent Interactions in HCOOH⋯Benzene Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2729**

Table 4. Binding Energies and Mean Errors (in kcal/mol) for DFT-D[a]

method	A	B	C	D	E	MSE[b]	MUE[c]
best estimate[d]	3.15	3.30	4.79	3.23	1.62		
M06-D	3.25	3.71	5.31	3.37	1.50	0.21	0.26
BLYP-D	3.52	3.58	5.22	3.61	1.38	0.24	0.34
M06-L-D	3.31	3.79	5.42	3.46	1.42	0.26	0.34
ωB97X-D	3.63	3.80	5.14	3.63	1.39	0.30	0.39
BMK-D	3.39	3.70	5.46	3.73	1.01	0.24	0.48
PBEsol-D	3.68	3.72	5.77	3.69	1.68	0.49	0.49
BP86-D	3.67	3.77	5.63	3.78	1.06	0.36	0.59
B3LYP-D	3.92	4.02	5.73	4.02	1.68	0.66	0.66
SOGGA-D	3.86	3.90	6.03	3.88	1.79	0.67	0.67
PBE-D	3.93	4.04	5.94	4.04	1.79	0.73	0.73
PBE0-D	4.06	4.16	6.08	4.16	1.79	0.83	0.83
TPSS-D	4.25	4.33	6.19	4.33	1.64	0.93	0.93

[a] Tested with the 6-311+G(2df,2p) basis set and MP2/6-311+G(2df,2p) geometries. [b] MSE denotes mean signed error (same as mean deviation). [c] MUE denotes mean unsigned error (same as mean absolute deviation, MAD). [d] Taken from the CCSD(T)/CBS results of Table 1.

From Table 1, we can see that the small-basis CCSD(T) corrections contribute −0.06 to −0.65 kcal/mol to the final dissociation energies. Among the four π-hydrogen bonded complexes, complex C has the largest binding energy; the hydrogen donor in complex C is the −OH group of formic acid, which is a stronger hydrogen bond donor than the C−H hydrogen donor in the other three π-hydrogen bonded complexes (A, B, and D). Complex E is the weakest HCOOH⋯benzene dimer.

The largest binding energy in HCOOH⋯benzene dimers is 4.79 kcal/mol (complex C), which is about 0.2 kcal/mol less than the binding energy in water dimer (5.02 kcal/mol).[67] The binding strength in complex C is stronger than those between benzene and first- and second-row hydrides (among them HF⋯benzene has the largest binding energy of 4.3 kcal/mol, and H_2O⋯benzene has a binding energy of 3.2 kcal/mol).[15] According to Imai et al. report,[16] the largest binding energy in $HCONH_2$⋯benzene is 3.0 kcal/mol, which is about 1.8 kcal/mol smaller than the binding energy in complex C; this result agrees with the fact that HCOOH is a much stronger hydrogen bond donor than $HCONH_2$.

3.3. Performance of Local Density Functionals. Table 2 gives the test results for local density functionals; we use the word "local" in the analogy as Becke,[68] so that it includes spin density gradient and spin kinetic energy density (as well as local spin density) but excludes Hartree−Fock exchange.

Among the tested local functionals, only LSDA and VSXC overestimate the binding energies, as shown by their positive MSEs. Other functionals underestimate the binding energies. Most of the tested local density functionals give a bad performance with the major exception being the M06-L, SOGGA, and PBEsol functionals, which gives a mean unsigned error (MUE) of 0.40, 0.57, and 0.76 kcal/mol, respectively. Other local functionals have MUEs greater than 1 kcal/mol. It is not surprising that M06-L give a good performance for noncovalent interactions in HCOOH⋯benzene complexes, since the training set of the M06-L functional contains some noncovalent interaction data. The surprisingly good performance of the nonempirical SOGGA and PBEsol functionals is likely due to their diminished gradient dependence[41] as compared to PBE. As shown in Table 2, LSDA overestimate the binding strength, whereas PBE underestimates them. In PBE, the second-order gradient expansion for exchange is not respected, whereas in SOGGA and PBEsol, the gradient expansion for exchange has been restored,[40,41] and their exchange enhancement factors are between LSDA and PBE, so both SOGGA and PBEsol perform better than PBE. At the high reduced gradient region, which is important for noncovalent interactions, SOGGA diminishes more gradient correction than PBEsol,[41] so SOGGA gives slightly stronger binding than PBEsol. Recent studies[40,41,69,70] showed that PBEsol and SOGGA give a very much better performance for lattice constant predictions than PBE and LSDA, but M06-L is less accurate for lattice constants than PBEsol and SOGGA. PBEsol has also been shown to give an improved description of stereoelectronic effects in hydrocarbons.[61]

3.4. Performance of Hybrid Density Functionals. Table 3 presents test results for hybrid functionals.

Among the tested hybrid functionals, PWB6K and M05-2X functionals give a good performance, as shown by their small MUEs (0.16 and 0.26 kcal/mol), followed by M06-2X. The best performing hybrid GGA is PBE0, which gives an MUE of 1.49 kcal/mol.

Hybrid functionals have a component of nonlocal Hartree−Fock exchange, thus their computational cost is much higher than the local functionals. In particular, for the HCOOH⋯benzene complex, the best local functional, M06-L, is about six times faster than the best hybrid functional, PWB6K. PBEsol and SOGGA are even less expensive than M06-L, and they are about eight times faster than PWB6K.

Table 5. M06-L Binding Energies (kcal/mol) with Different Geometries and Basis Sets

method	A	B	C	D	E	MSE[a]	MUE[b]
best estimate	3.15	3.30	4.79	3.23	1.62		
M06-L/DIDZ//M06-L/DIDZ	3.06	3.24	4.65	3.05	1.34	−0.15	0.15
M06-L/6-311+G(2df,2p)//M06-L/6-311+G(2df,2p)	2.87	3.16	4.78	2.89	1.44	−0.19	0.19
M06-L/DIDZ//MP2/6-311+G(2df,2p)	2.91	3.17	4.49	2.98	1.30	−0.25	0.25
M06-L/DIDZ//M06-L/MIDI!	2.96	2.84	4.45	2.64	1.38	−0.36	0.36
M06-L/aug-cc-pVTZ//MP2/6-311+G(2df,2p)	2.67	2.99	4.74	2.80	1.00	−0.38	0.38
SOGGA/6-31+G(d)//SOGGA/6-31+G(d)	2.62	2.55	5.36	2.59	1.38	−0.32	0.54
SOGGA/6-311+G(2df,2p)//MP2/6-311+G(2df,2p)	2.47	2.38	4.56	2.53	1.29	−0.57	0.57
PBEsol/6-31+G(d)//PBEsol/6-31+G(d)	2.45	2.41	5.14	2.46	1.26	−0.47	0.61
PBEsol/6-311+G(2df,2p)//MP2/6-311+G(2df,2p)	2.29	2.20	4.30	2.35	1.18	−0.76	0.76

[a] MSE denotes mean signed error (same as mean deviation). [b] MUE denotes mean unsigned error (same as mean absolute deviation, MAD).

Figure 2. Intermolecular potential of the HCOOH···benzene complex. DFT calculations employed the 6-311+G(2df,2p) basis set without Cp corrections.

Figure 3. Intermolecular potential of the HCOOH···benzene complex. DFT-Ex and HF calculations employed the 6-311+G(2df,2p) basis set without Cp corrections. The PBE-Ex (PBE exchange-only) calculations were done without PBE correlation contributions using the PBE SCF density. The TPSS-Ex (TPSS exchange-only) calculations were done without TPSS correlation contributions using the TPSS SCF density.

3.5. Performance of DFT-D. The results for different DFT-D methods are shown in Table 4. Table 4 confirms the finding in previous studies[60−63] that empirical dispersion corrections improve the description of noncovalent interactions. All the DFT-D methods in Table 4 have MUEs less than 1.0 kcal/mol. Moreover, the MUE of BLYP-D is about 11 times smaller than that of the original BLYP, showing the effectiveness of the empirical dispersion corrections for the description of noncovalent interactions in HCOOH···benzene complexes. The second best GGA-D method is PBEsol-D. SOGGA-D uses the same empirical dispersion corrections as PBEsol-D, but SOGGA-D gives a slightly larger MUE than PBEsol-D.

3.6. Effects of Geometries and Basis Sets. In previous sections, we based our discussions on single-point energies calculated with the MP2/6-311+G(2df,2p) geometries. In this section, we compare results with the geometries optimized at the M06-L/MIDI!, M06-L/6-31+G(d,p), SOGGA/6-31+G(d), PBEsol/6-31+G(d), and M06-L/6-311+G(2df,2p) levels of theories, where MIDI![71] is a well-balanced and economical double-ζ basis set for geometry optimization. The results are shown in Table 5.

As shown in Table 5, for a given functional, the effect of geometries and basis sets for the binding energies is small, and they agree with each other within 0.8 kcal/mol. M06-L/6-31+G(d,p) gives a smallest MUE (0.15 kcal/mol), and M06-L/6-31+G(d,p)//M06-L/MIDI! has an MUE of 0.36 kcal/mol. The good performance with the double-ζ quality basis set is likely due to a well-known basis set error that diminishes the underbinding error of an exchange-correlation functional, as shown recently by Ruzsinszky et al.[72] Similarly, PBEsol/6-31+G(d) gives a better performance than PBEsol/6-311+G(2df,2p).

3.7. Counterpoise Corrected DFT Results. Table 6 presents the Cp corrected binding energies and mean errors. After removing the BSSEs, the best two performers in Table
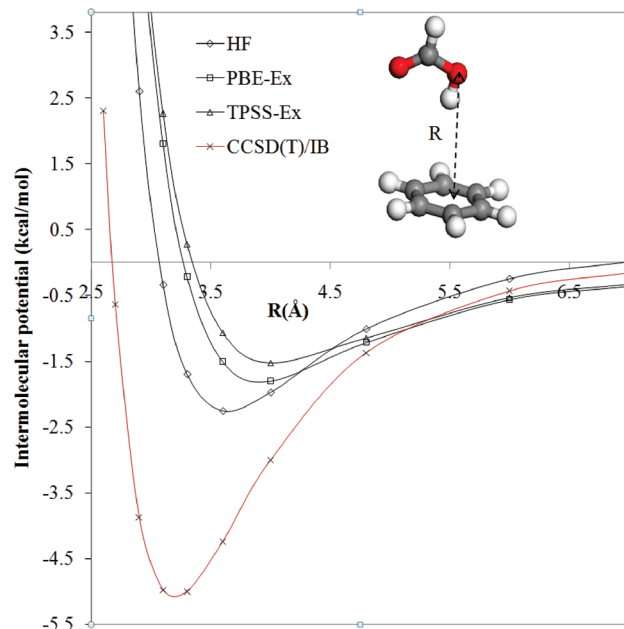
6 are BLYP-D and PBEsol-D, with an MUE of 0.15 kcal/mol. Among the ten best performing functionals, only M05-2X and M06-2X do not have empirical dispersion corrections, and all other eight functionals are DFT-D methods. This result again shows the effectiveness of empirical dispersion corrections and also confirms the finding in Section 3.5.

3.8. Intermolecular Potential of Complex C. We have calculated the intermolecular potential of complex C with the M06-L/6-311+G(2df,2p), PBEsol/6-311+G(2df,2p), SOGGA/6-311+G(2df,2p), and CCSD(T)/IB level of theories, and the intermolecular distance is defined as the distance between the oxygen atom of hydrogen donor in HCOOH and the center of mass of the benzene molecule. The intermolecular potentials are shown in Figure 2.

As shown in Figure 2, the M06-L, PBEsol, and SOGGA intermolecular potential curves show reasonable agreement with that of CCSD(T)/IB. Note that the more accurate CCSD(T)/CBS minimum is above the CCSD(T)/IB minimum by 0.25 kcal/mol, thus shows better agreement with the DFT results. Figure 2 also shows that, even at 5 Å apart, the interaction energy is greater than 1 kcal/mol, which is about twice of the binding energy of neon···benzene van der Waals complex. This result indicates that the interaction between HCOOH and benzene is not just due to dispersion interactions, and the electrostatic contributions play an important role in the HCOOH···benzene complexes.

Results in Figure 2 confirm that GGAs or meta-GGAs, even without empirical dispersion correction, might provide a reasonable description of the noncovalent interactions arising from overlapped nonbonded density. Nevertheless, GGAs or meta-GGAs cannot describe the long-range disper-

Table 6. Counterpoise Corrected Binding Energies and Mean Errors (in kcal/mol)

method[a]	A	B	C	D	E	MSE[b]	MUE[c]
best estimate[d]	3.15	3.30	4.79	3.23	1.62		
BLYP-D	3.10	3.16	4.73	3.19	1.19	−0.15	0.15
PBEsol-D	3.17	3.19	5.16	3.18	1.43	0.01	0.15
ωB97X-D	3.19	3.35	4.61	3.20	1.15	−0.12	0.15
SOGGA-D	3.34	3.37	5.40	3.35	1.55	0.18	0.21
M05-2X	2.85	3.11	4.87	2.96	1.39	−0.18	0.21
M06-2X	3.10	3.58	5.17	3.33	1.35	0.09	0.22
M06-L-D	2.81	3.25	4.85	2.95	1.12	−0.22	0.25
BP86-D	3.22	3.31	5.10	3.33	0.85	−0.06	0.25
BMK-D	3.02	3.31	5.04	3.37	0.80	−0.11	0.27
M06-D	2.74	3.17	4.74	2.87	1.21	−0.27	0.27
B3LYP-D	3.48	3.58	5.22	3.59	1.48	0.25	0.31
PBE-D	3.47	3.57	5.38	3.58	1.58	0.30	0.31
PBE0-D	3.56	3.67	5.49	3.67	1.57	0.37	0.39
TPSS-D	3.72	3.80	5.57	3.80	1.40	0.44	0.53
M06-HF	3.62	4.10	5.68	3.79	1.37	0.50	0.60
PWB6K	2.42	2.51	4.08	2.63	1.41	−0.61	0.61
M06-L/DIDZ//MP2/6-311+G(2df,2p)	2.42	2.65	3.96	2.47	1.02	−0.72	0.72
M06-L/DIDZ/M06-L/MIDI!	2.47	2.29	3.92	2.15	1.09	−0.83	0.83
M06-L	2.10	2.44	4.03	2.21	0.89	−0.88	0.88
M06	2.03	2.36	3.92	2.12	0.97	−0.94	0.94
M06-L/aug-cc-pVTZ//MP2/6-311+G(2df,2p)	2.02	2.34	4.01	2.17	0.71	−0.97	0.97
LSDA	4.02	4.29	6.51	4.26	2.24	1.05	1.05
SOGGA	1.94	1.85	3.93	2.01	1.05	−1.06	1.06
BHandH	4.18	4.46	6.43	4.41	2.49	1.18	1.18
PBEsol	1.78	1.67	3.68	1.83	0.94	−1.24	1.24
PW6B95	1.73	1.73	3.26	1.91	1.01	−1.29	1.29
M05	1.37	1.33	2.87	1.49	1.34	−1.54	1.54
B97-1	1.12	0.89	2.47	1.10	0.95	−1.91	1.91
PBE0	1.09	0.85	2.61	1.06	0.75	−1.95	1.95
PBE	0.81	0.55	2.30	0.79	0.70	−2.19	2.19
BMK	0.72	0.69	2.37	0.96	0.04	−2.26	2.26
MPW1K	0.81	0.51	2.24	0.74	0.47	−2.26	2.26
B98	0.74	0.45	2.02	0.68	0.72	−2.30	2.30
MPW3LYP	0.65	0.34	1.89	0.60	0.93	−2.34	2.34
BHandHLYP	0.63	0.32	1.87	0.58	0.76	−2.39	2.39
τ-HCTHh	0.55	0.28	1.93	0.50	0.40	−2.49	2.49
mPW1PW	0.41	0.07	1.79	0.33	0.31	−2.64	2.64
X3LYP	0.30	−0.04	1.52	0.25	0.64	−2.69	2.69
TPSSh	0.34	−0.05	1.64	0.25	0.29	−2.72	2.72
TPSS1KCIS	0.32	−0.07	1.53	0.23	0.43	−2.73	2.73
TPSS	0.18	−0.23	1.46	0.09	0.23	−2.87	2.87
mPWPW91	−0.10	−0.52	1.19	−0.21	0.08	−3.13	3.13
B3LYP	−0.23	−0.64	0.90	−0.32	0.26	−3.23	3.23
B97-2	−0.22	−0.65	0.95	−0.35	0.07	−3.26	3.26
HCTH	−0.52	−1.05	0.44	−0.70	0.74	−3.44	3.44
τ-HCTH	−0.48	−1.06	0.49	−0.70	0.17	−3.53	3.53
BP86	−0.50	−0.92	0.78	−0.57	−0.38	−3.54	3.54
BLYP	−1.15	−1.68	−0.21	−1.27	−0.21	−4.12	4.12
O3LYP	−2.10	−2.89	−1.28	−2.37	−0.49	−5.05	5.05
OLYP	−2.98	−3.89	−2.34	−3.29	−0.88	−5.90	5.90
VSXC	8.76	11.26	12.25	10.35	4.01	6.11	6.11
G96LYP	−3.30	−4.10	−2.53	−3.51	−1.95	−6.30	6.30

[a] If basis sets and geometries were not specified, the DFT calculations employed the 6-311+G(2df,2p) basis set and MP2/6-311+G(2df,2p) geometries. [b] MSE denotes mean signed error (same as mean deviation). [c] MUE denotes mean unsigned error (same as mean absolute deviation, MAD). [d] Taken from the CCSD(T)/CBS results of Table 1.

sion interactions in nonoverlapped densities. They give an exponential decay instead of the $-C_6R^{-6}$ behavior for the long-range intermolecular potential.

Table 1 showed that HF theory severely underestimates the interaction energies in the HCOOH···benzene complexes. An interesting observation from Tables 2 and 3 is that some global hybrid functionals (e.g., PBE0 and TPSSh) improve slightly upon corresponding local functionals (e.g., PBE and TPSS). In order to understand the origin of this improvement, we calculated the intermolecular potential with HF theory, PBE-Ex (PBE exchange-only calculation using

the PBE SCF densities), and TPSS-Ex (TPSS exchange-only calculation using TPSS SCF densities). The potential curves are plotted in Figure 3. As shown in Figure 3, for the intermolecular distance at 3−4 Å, HF theory gives more attractive interaction than PBE and TPSS exchange functionals. Thus, replacing certain amount of PBE or TPSS exchange with the HF exchange (as done in PBE0 or TPSSh) results in more attractive interactions at the 3−4 Å region. This explains why PBE0 and TPSSh slightly improve upon the corresponding local PBE and TPSS functionals. Interestingly, PBE and TPSS exchange functionals give more

2732 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Zhao et al.

attractive interactions than HF theory at large intermolecular distance (>4.5 Å), as shown in Figure 3.

4. Conclusions

In the present study, we have employed the MP2 and CCS(T) methods to study structures and interactions in the HCOOH···benzene complexes, and we have developed a benchmark data for binding energies in five noncovalent HCOOH···benzene dimers. The benchmark data were determined as the sum of the complete-basis-set limit of MP2 energies and a CCSD(T) correction term evaluated with the aug-cc-pVDZ basis set. Final binding energies (kcal/mol) are in the range of 1.6−4.8 kcal/mol, and they were used as reference data to test density functionals in the literature. Based on these results, we draw the following conclusions:

1) The small-basis CCSD(T) contributions to the final best estimates of the binding energies are small, with magnitudes in the range of −0.06 to −0.65 kcal/mol.

2) The calculated benchmark data and intermolecular potential are useful for the parametrizations of new force fields and coarse-grained models for chemical species such as the acrylic polymers.

3) The largest binding energy in HCOOH···benzene dimers is 4.8 kcal/mol (complex C), and this binding strength is weaker than that of water dimer but stronger than noncovalent interactions between benzene and first- and second-row hydrides.

4) Among the tested local density functionals without empirical dispersion corrections, M06-L/6-31+G(d,p) gives an MUE of only 0.15 kcal/mol. PBEsol and SOGGA also show promising performance.

5) Empirical dispersion corrections greatly improve the performance of DFT for the descriptions of noncovalent interactions in HCOOH···benzene dimers.

Supporting Information Available: Cartesian coordinates of the HCOOH···benzene dimers optimized at the MP2/6-311+G(2df,2p) level. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Atwood, H.; Hamada, F.; Robinson, D. K.; Orr, G. W.; Vincent, R. L. *Nature* **1991**, *349*, 603.

(2) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem. Int. Ed.* **2003**, *42*, 1210.

(3) Lindeman, S. V.; Kosynkin, D.; kochi, J. K. *J. Am. Chem. Soc.* **1998**, *120*, 13268.

(4) Maity, S.; Patwari, G. N. *J. Phys. Chem. A* **2009**, *113*, 1760.

(5) Monkman, A. P.; Pålsson, L.-O.; Higgins, R. W. T.; Wang, C.; Bryce, M. R.; Batsanov, A. S.; Howard, J. A. K. *J. Am. Chem. Soc.* **2002**, *124*, 6049.

(6) Desiraju, G. R. *Chem. Commun.* **2005**, 2995.

(7) Zhao, R.; Matsumoto, S.; Akazome, M.; Ogura, K. *Tetrahedron* **2002**, *58*, 10233.

(8) Feller, D. *J. Phys. Chem. A* **1999**, *103*, 7558.

(9) Gruenloh, C. J.; Carney, J. R.; Arrington, C. A.; Zwier, T. S.; Fredericks, S. Y.; Jordan, K. D. *Science* **1997**, *276*, 1678.

(10) Tarakeshwar, P.; Choi, H. S.; Kim, K. S. *J. Am. Chem. Soc.* **2001**, *123*, 3323.

(11) Tarakeshwar, P.; Kim, K. S.; Brutschy, B. *J. Chem. Phys.* **2000**, *112*, 1769.

(12) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 11450.

(13) Baron, M.; Kowalewski, V. J. *J. Phys. Chem. A* **2006**, *110*, 7122.

(14) Slipchenko, L. V.; Gordon, M. S. *J. Phys. Chem. A* **2009**, *113*, 2092.

(15) Crittenden, D. L. *J. Phys. Chem. A* **2009**, *113*, 1663.

(16) Imai, Y. N.; Inoue, Y.; Nakanishi, I.; Kitaura, K. *J. Comput. Chem.* **2009**, *30*, 2267.

(17) Ottiger, P.; Pfaffen, C.; Leist, R.; Leutwyler, S.; Bachorz, R. A.; Klopper, W. *J. Phys. Chem. B* **2009**, *113*, 2937.

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

(19) Paldus, J. The beginning of coupled-cluster theory: an eyewitness account. In *Theory and Application of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 115.

(20) Bartlett, R. J. How and why coupled-cluster theory became the pre-eminent method in an ab initio quantum chemistry. In *Theory and Application of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 1191.

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

(22) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45.

(23) Zhao, Y.; Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 19046.

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

(25) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(26) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418.

(27) Mahmud, S.; Davidson, E. R. S. S. *Surf. Sci.* **1995**, *322*, 342.

(28) Perdew, J. P.; Schmidt, K. In *Density Functional Theory and Its Applications to Materials*; Van-Doren, V., Alsenoy, C. V., Geerlings., P., Eds.; American Institute of Physics: New York, 2001; p 1.

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

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

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

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

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

(34) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

(35) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433.

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

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

(38) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

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

(40) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. *Phys. Rev. Lett.* **2008**, *100*, 136406.

(41) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2008**, *128*, 184109.

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

(43) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.

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

(45) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.

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

(47) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.

(48) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.

(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Wong, M. W.; Foresman, J. B.; Robb, M. A.; Head-Gordon, M.; Replogle, E. S.; Gomperts, R.; Andres, J. L.; Raghavachari, K.; Binkley, J. S.; Gonzalez, C.; Martin, R. L.; Fox, D. J.; Defrees, D. J.; Baker, J.; Stewart, J. J. P.; Pople, J. A. *Gaussian 92/DFT; Revision F.2 ed.*; Gaussian, Inc.: Pittsburgh, PA, 1993.

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

(51) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(52) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319.

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

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

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

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

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

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

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

(60) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(61) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P.; Grimme, S. *J. Chem. Theory Comput.* **2008**, *4*, 888.

(62) Chai, J.-D.; Martin, H.-G. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(63) Karton, A.; Tarnopolsky, A.; Lamre, J.-F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2009**, *112*, 12868.

(64) Gordon, M. S.; Schmidt, M. W. In *Theory and Applications of Computational Chemistry: the first 40 years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 1167.

(65) Bylaska, E. J.; Jong, W. A. d.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Aprà, E.; Windus, T. L.; Hammond, J.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; Lenthe, J. v.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*; Version 5.1 ed.; Pacific Northwest National Laboratory: Richland, WA, U.S.A., 2007.

(66) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian09; Revision A.02 ed.*; Gaussian, Inc.: Pittsburgh, PA, 2009.

(67) Tschumpera, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; IIIb, H. F. S.; Quack, M. *J. Chem. Phys.* **2002**, *116*, 690.

(68) Becke, A. D. *J. Chem. Phys.* **1998**, *109*, 2092.

(69) Haas, P.; Tran, F.; Blaha, P. *Phys. Rev. B* **2009**, *79*, 85104.

(70) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A.; Philipsen, P. H. T.; Lebègue, S.; Paier, J.; Vydrov, O. A.; Ángyán, J. G. *Phys. Rev. B* **2009**, *79*, 155107.

(71) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93*, 281.

(72) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I. *J. Phys. Chem. A* **2005**, *109*, 11015.

JCTC Journal of Chemical Theory and Computation

An Automated and Systematic Transition Structure Explorer in Large Flexible Molecular Systems Based on Combined Global Reaction Route Mapping and Microiteration Methods

Satoshi Maeda,[†,‡] Koichi Ohno,*[,‡] and Keiji Morokuma*[,†,§]

Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, Toyota Physical and Chemical Research Institute, Nagakute, Aichi 480-1192, Japan, and Fukui Institute for Fundamental Chemistry, Kyoto University, Kyoto 606-8103, Japan

Abstract: The global reaction route mapping (GRRM) method enabled an automated and a systematic search for routes of chemical reactions on a potential energy surface based on the anharmonic downward distortion following (ADDF) approach [*Chem. Phys. Lett.* **2004**, *384*, 277]. On the other hand, the microiteration technique [*Mol. Phys.* **2006,** *104*, 701] has been developed for full optimizations of transition state (TS) structures for reactions/transformations in large flexible molecular systems and successfully used in ONIOM(QM:MM) calculations. In the present paper, combining the GRRM method with the microiteration technique, we developed a microiteration-ADDF (μ-ADDF) method for automated and systematic TS exploration of large flexible molecular systems. We showed that the method works well with two test systems, $(H_2CO)(H_2O)_{100}$ and $Si_6(C_{12}H_{17})_6$, in the ONIOM(QM:MM) framework. It is noted that the present μ-ADDF method can be used for pure quantum mechanics (QM) or molecular mechanics (MM) systems (without ONIOM) and has been tested successfully in $C_6H_{10}O$ pure QM calculations.

I. Introduction

Determination of transition state (TS) structures, first-order saddle points on a potential energy surface (PES), is essential in theoretical studies of chemical reaction mechanisms.[1,2] Quantum mechanics/molecular mechanics (QM/MM)[3−5] and the more general ONIOM[6−8] approaches made applications of accurate quantum methods, such as ab initio and density functional methods, possible to very large molecular systems including biological molecules. For discussing reaction mechanisms in such systems, a search is required for TS structures of chemical bond rearrangements in a given reaction center region surrounded by large substituents or environmental systems such as protein.

The Newton−Raphson method or the more sophisticated augmented Hessian approaches, such as the Berny optimization algorithm[9] and the eigenvector following[10] (EF) and the rational function optimization[11] (RFO) methods, are employed in TS optimizers. Efficiency of these methods strongly depends on the overall algorithm including the selection of coordinate system,[12−14] the Hessian update scheme,[15−18] the use of the trust radius for limiting step size,[19−21] and so on.[1] Obviously, standard Hessian-based approaches are not applicable to very large molecular systems without special care because of the following two difficulties: (1) quadratic and cubic increase of Hessian storage and diagonalization costs, respectively, with respect to the numbers of atoms N, and (2) selecting initial geometries with only one imaginary frequency eigenvector parallel to desired reaction routes from extensive coordinate space of such systems.

Both of the above two difficulties can be greatly circumvented by use of the microiteration technique.[22−28] In a geometry optimization algorithm based on the microiteration

  * Corresponding authors. E-mail: ohnok@mail.tains.tohoku.ac.jp (K.O.) and morokuma@emory.edu (K.M.).
  † Emory University.
  ‡ Toyota Physical and Chemical Research Institute.
  § Kyoto University.

technique in ONIOM or QM/MM method, optimizations of MM atom positions fixing all QM atoms (microiteration) are performed before every optimization step of QM atom positions (macroiteration). Since MM atom positions can be removed from optimization variables in macroiterations, this algorithm greatly simplifies the problem preparing initial structures with the correct Hessian structure. The difficulty concerning the size of a Hessian matrix is completely removed in microiterations because first-order optimization techniques, such as the conjugated gradient method, can be employed in microiterations. Although many more force calculations will be required in first-order methods than in second-order methods, which use (exact or updated) Hessian, this does not matter in microiterations since the cost of force calculations is trivial in MM calculations. In the microiteration method, the coupling of both the QM and MM parts is usually completely neglected in the QM Hessian calculation. Recently, a fully coupled optimization algorithm was proposed in which quadratic couplings between the QM and MM regions are taken into account explicitly in macroiterations,[29] and it dramatically reduced the number of macroiteration steps. Although the first difficulty again arises in the fully coupled scheme since diagonalization of full Hessian matrices, including QM/QM, QM/MM, and MM/MM terms, is required in every macroiteration to take the quadratic couplings into account, its cost can be reduced to O(N) scaling by using sophisticated numerical techniques, such as the iterative Davidson diagonalization[30] algorithm and the fast multipole[31,32] method.[29] This fully coupled microiteration scheme implemented in a development version of Gaussian and recently released Gaussian 09[33] programs enabled efficient full optimization of ONIOM(QM:MM) TSs and intermediates of complex systems, such as the entire mechanism consisting of multiple steps for the nonheme iron enzyme isopenicillin N synthase containing as many as 5 368 atoms.[34]

Augmented Hessian methods, when used in general TS optimizers, maximize the energy in the direction of the only imaginary frequency eigenvector. Toward this goal, a user has to prepare a good initial guess, which already has a negative eigenvalue (or a lowest positive eigenvalue, which has eigenvector parallel to the negative eigenvalue direction at a final TS structure). Therefore, a method that does not require any initial guess of TS structures is desired to be combined with the microiteration technique to find all of the important TS structures for a chemical reaction. Here, two situations can be considered: (1) both reactant and product structures are known beforehand, and (2) only a reactant structure is available. In the first case, there are many good so-called double-ended methods,[35−43] and some of them can be applied to multistep cases without knowledge of all intermediates. One of them, the synchronous transit-guided quasi-Newton method[36] has already been used in combination with the microiteration technique in Gaussian programs, although it cannot be employed for multistep reactions unless one prepares a series of all intermediate structures beforehand. In the second case, one has to use one-point methods which gives all of the intermediates and the TSs automatically from only one input structure. There are three such

approaches which have actually been applied to automated global mapping of ab initio PESs: (1) the gradient extremal following[44−53] (GEF), (2) the reduced gradient following[54−60] (RGF), and (3) the anharmonic downward distortion following[61−63] (ADDF). The former two have been applied to a system composed of only four atoms.[53,54] On the other hand, we proposed the global reaction route mapping (GRRM) method on the basis of the ADDF approach,[61−63] and the GRRM method enabled automated and systematic global mapping on PESs of given chemical formula in many small systems.[64−69] Hence, the GRRM should be the method of choice to be combined with the microiteration technique for exploring PESs of reaction center variables in large flexible environments.

II. Theory

The GRRM Method. If reaction routes can be followed in both uphill and downhill directions, then GRRM can be performed as follows: (1) an equilibrium (EQ) structure is optimized starting from an arbitrary input structure; (2) the optimized EQ is added to a list of EQ structures (EQ-list); (3) all entrances of reaction pathways are searched at one of EQs in the EQ-list; (4) each pathway is followed in an uphill direction toward a TS or a dissociation channel (DC); (5) starting from each TS obtained, an EQ is searched by following the reaction pathway in a downhill direction; (6) new EQs are added to the EQ-list; (7) return to (3) if there is any EQ in the EQ-list to which the procedures (3−6) have not yet been applied; and (8) exit from the cycle if not. A completion of this procedure is expected to give a global reaction route map including all EQs and TSs for a given chemical formula. Here, downhill walks from TSs can be performed by using conventional intrinsic reaction coordinate (IRC) following techniques.[70−73] Unfortunately, there is no technique which is mathematically guaranteed to find out all entrances of reaction pathways from an EQ structure. Recently, we suggested a principle that all reaction pathways can be found as ADD maxima around EQ points.[61−63] Although this is not a mathematical theorem either, we have shown that it worked very well in many previous applications.[64−69] Hence, we proposed the GRRM method using both uphill (ADDF)[61−63] and downhill (IRC)[70−73] methods.

Some approximate treatments of the GRRM method can be considered in which the procedures (3−6) are applied to only selected EQs, which are important in a specific problem. For instance, they are applied to only stable EQs or TSs with less than a given total energy, or they are applied to only EQs possessing a specified bonding pattern. All of these treatments are available in the latest version of the GRRM program.

The ADDF Approach. Many typical potential curves show a common feature that potential energy always becomes lower than the harmonic potential defined at the bottom of the curves in directions leading to DCs and TSs. From this feature, we proposed a principle that reaction channels can be found by following ADD maxima starting from an EQ structure on a PES.[61−63] In other words, existence of a flat region of PES or another EQ should have a certain influence upon PES around the starting EQ, and such indications can

be detected as ADD maxima at the starting EQ. In many previous applications, the GRRM method based on the ADDF approach has found many unknown as well as (almost) all known reaction channels automatically.[61−69] The ADD maxima can be detected in multidimension by using the scaled hypersphere search (SHS) technique as explained below.

There are many local ADD maxima around a starting EQ, and in the full-ADDF (*f*ADDF) approach all of them are followed.[61−69] We suggested that larger ADDs are related to lower barrier pathways leading to lower energy EQs,[74] by combining the Bell−Evans−Polanyi (BEP) principle[75,76] with the above ADD principle. The BEP principle tells that the deeper minimum of the product side is related to the lower barrier, whereas the ADD principle says that the deeper minimum (strong potential lowering interaction) of the product side is related to the larger ADD of the reactant side. It follows that the larger ADD is related to the lower barrier leading to the deeper minimum, and high barrier pathways can be omitted by comparisons of magnitude of ADDs at the starting EQ without following them. In the large-ADDF[74] (*l*ADDF) approach, one follows only large ADDF pathways avoiding those suggesting high barriers; this approach has been applied to H-bond cluster systems[74,77−79] and conformation samplings.[80,81]

The SHS Technique. ADD maxima are searched by comparisons between harmonic and real (QM, MM, or QM/MM) energies on an isoenergy hypersurface of harmonic potential. Such a hypersurface is a hyperellipsoid centered at the starting EQ and can be converted to a simple hypersphere if the scaled normal coordinates $q_i = \lambda_i^{1/2}Q_i$ are employed, where λ_i is the eigenvalue of Q_i. In the SHS technique, ADD maxima can be detected as energy minima on the scaled hyperspheres, and they are traced by expanding the hypersphere radius. The method to find all or *n*-lowest ADDs on a given scaled hypersphere was discussed in the previous papers[62,74] and will not be repeated. Here, use of the harmonic reference is essential in the ADDF by the SHS method, and how to define a good harmonic reference for reaction center variables was a key point of the present development.

An angle coordinate $(\theta_1, \theta_2, ..., \theta_{f-1})$ is used as variables in the energy minimizations on the scaled hyperspheres with a given radius r (in hartree$^{1/2}$) as in the polar coordinate interpolation.[82] In the latest implementation, $(\theta_1, \theta_2, ..., \theta_{f-1})$ are defined using a set of orthogonal vectors $(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{f-1}, \mathbf{v})$ as $f = 3N - 6$ axes. At first, a Cartesian Hessian (exact or updated) is converted to the one in the scaled normal coordinate. This is a simple coordinate rotation and scaling. Here, $\mathbf{v}$ is a unit vector pointing to the present position from the origin (starting EQ), and projections of $\mathbf{v}$ are eliminated from the Hessian by using the projection method.[83] Then, diagonalization of the projected Hessian gives $f - 1$ eigenvectors $\mathbf{u}_i$ with nonzero eigenvalues ε_i. Based on an expansion at the present position with $\theta_i = 0$ for all i, u_i, and v at other points on the sphere can be written as

$$
\begin{aligned}
u_i &= r\sin\theta_i \prod_{j}^{i-1}\cos\theta_j \\
v &= r\prod_{j}^{f-1}\cos\theta_j
\end{aligned}
\tag{1}
$$

From eq 1, $\partial u_i/\partial\theta_i = r$, $\partial u_i/\partial\theta_j = 0$ $(i \neq j)$, $\partial^2 u_i/\partial\theta_j\partial\theta_k = 0$, $\partial v/\partial\theta_i = 0$, $\partial^2 v/\partial\theta_i^2 = r$, $\partial^2 v/\partial\theta_i\partial\theta_j = 0$ $(i \neq j)$ can be obtained at the present position with $\theta_i = 0$ for all i. Then, these conditions and $\partial^2 E/\partial u_i\partial u_j = 0$ $(i \neq j)$ give

$$
\begin{aligned}
\frac{\partial E}{\partial\theta_i} &= r\frac{\partial E}{\partial u_i} \\
\frac{\partial^2 E}{\partial\theta_i^2} &= r^2\varepsilon_i + r\frac{\partial E}{\partial v}
\end{aligned}
\tag{2}
$$

Here, Hessian in terms of θ_i is already diagonal, i.e., $\partial^2 E/\partial\theta_i\partial\theta_j = 0$ $(i \neq j)$, without further diagonalization, and $\partial E/\partial\theta_i$ and $\partial^2 E/\partial\theta_i^2$ are used as gradient components and Hessian eigenvalues in an optimizer in the GRRM program based on the RFO and trust radius methods.

Double-Ended Methods in the GRRM Program. We also have a double-ended method in the GRRM program, which requires two initial structures but is much faster than the one-point method in general, since the area to be searched can be limited to a small region between the two structures.[42,43] In the double-ended ADDF (*d*ADDF) method, energy minima on the scaled hyperspheres are followed in the reverse direction from a point on a very large hypersphere to the sphere center with reducing hypersphere radius. If there exists only one TS between the two structures, the trace of ADD maximum points is expected to pass through the TS region, same as the original sphere expansion ADDF algorithm. Hence, the *d*ADDF can be used as a guide to a TS region between two adjacent EQs, which we call *d*ADDF-guided TS search (*d*ADDF-TS) approach.[42] On the other hand, if there exists many EQ structures between the two chosen starting structures, a harmonic reference defined at a sphere center is no longer meaningful at a very long distance from the sphere center. In such a case, the trace of ADD maximum points may not pass through TS regions. However, each EQ (minima in *f* dimension) are still on the traces of minima on hyperspheres (minima in $f - 1$ dimension) followed by the *d*ADDF, and the *d*ADDF will give many EQs (intermediates) along its traces, where we call this procedure as *d*ADDF-guided EQ search (*d*ADDF-EQ) approach.[43] It is not guaranteed that all intermediates are obtained by only one application of the *d*ADDF-EQ to highly multistep and/or highly curved pathways, and one has to add some EQs by further applications of the *d*ADDF-EQ to pairs of EQs obtained in the initial application.[43] Finally, one can find a set of TSs connecting the intermediates by applying *d*ADDF-TS method to adjacent pairs in a set of EQs prepared by the *d*ADDF-EQ.

Effective Gradient and Hessian in the μ-ADDF Algorithm. We assume that there are p $(= 3N - 6)$ reaction-center variables x (coordinates directly involved in the reaction) and q $(= 3M)$ nonreaction-center variables y in a system composed of $N + M$ atoms. A second-order potential function for this system can be written as

Transition Structure Explorer in Large Systems

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2737**

$$V = \varepsilon_0 + \sum_i^p g_i^R x_i + \frac{1}{2} \sum_i^p \sum_j^p h_{ij}^R x_i x_j +$$
$$\frac{1}{2} \sum_i^p \sum_j^q h_{ij}^C x_i y_j + \frac{1}{2} \sum_i^q \sum_j^p h_{ij}^C y_i x_j +$$
$$\frac{1}{2} \sum_i^q \sum_j^q h_{ij}^N y_i y_j \quad (3)$$

where g^R are gradients for reaction-center variables, h^R, h^N, and h^C are Hessian matrix elements for reaction-center and nonreaction-center variables and their couplings, respectively. Here, all gradient components for nonreaction-center variables g^N are zero because they are already optimized in microiteration. In macroiterations, PESs of reaction-center variables fulfilling the following condition for nonreaction-center variables is considered

$$\frac{\partial V}{\partial y_n} = \sum_j^p h_{nj}^C x_j + \sum_j^q h_{nj}^N y_j = 0 \quad (4)$$

We rewrite eq 4 as the following simultaneous equations:

$$\sum_j^q h_{nj}^N y_j = - \sum_j^p h_{nj}^C x_j \quad (5)$$

Solutions of eq 5 for nonreaction-center variables are

$$y_m = - \sum_n^q \sum_j^p \gamma_{nm}^N h_{nj}^C x_j \quad (6)$$

where γ^N are elements of inverse Hessian matrix for nonreaction-center variables. On the other hand, the following condition can be obtained when summation over n is taken for eq 5 multiplied by y_n.

$$\sum_n^q \sum_j^q y_n h_{nj}^N y_j = - \sum_n^q \sum_j^p y_n h_{nj}^C x_j \quad (7)$$

From eq 7, the fifth and sixth terms in eq 3 are canceled out. Then, substitution of eq 6 into eq 3 without the fifth and sixth terms gives an effective second-order potential for reaction-center variables as

$$V = \varepsilon_0 + \sum_i^p g_i^R x_i + \frac{1}{2} \sum_i^p \sum_j^p h_{ij}^R x_i x_j -$$
$$\frac{1}{2} \sum_i^p \sum_m^q \sum_n^q \sum_j^q h_{im}^C \gamma_{nm}^N h_{nj}^C x_i x_j \quad (8)$$

This tells that the effective gradients for reaction-center variables are identical to the original gradients without any correction, and effective Hessian matrix elements for reaction-center variables can be obtained as

$$\frac{\partial^2 V}{\partial x_i \partial x_j} = h_{ij}^R - \sum_m^q \sum_n^q h_{im}^C \gamma_{nm}^N h_{nj}^C \quad (9)$$

Implementation of the Microiteration Technique in the GRRM Program. We implemented an interface with ONIOM in a previous version of GRRM program.[80] However, the *l*ADDF algorithm there worked only in conformation samplings of flexible parts in large systems because the ADDs related to conformation changes (with only a few kJ/mol barriers) are much larger than the ones for chemical bond reorganization (typically with more than 10 kJ/mol barriers). Although, in principle, the *f*ADDF can find out all

reaction routes, including chemical bond rearrangement pathways, it is not applicable in practice to very large systems. By combining with the microiteration technique, the *f*ADDF and *l*ADDF can be a practical automated TS explorer for chemical bond rearrangement reactions in large flexible systems. In the present implementation, all movements of reaction-center atoms were treated by the GRRM program as in the case without microiterations, where g^R and the (exact or updated) effective Hessian were employed for determining each displacement of atoms. Before every macroiterations (in minimizations on scaled hyperspheres, EQ and TS optimizations and IRC followings), positions of nonreaction-center atoms were optimized by using a microiteration code in the Gaussian programs, where opt = tight criterion was employed. All μ-ADDF algorithms, i.e., μ-*f*ADDF, μ-*l*ADDF, and μ-*d*ADDF, are now available in our GRRM program coupled with the Gaussian programs (only 03[84] and 09[33] versions).

It is noted that the microiteration technique is not necessarily limited to the ONIOM method. Division of the reaction-center and nonreaction-center variables can be made also within a pure QM system, a pure MM system, or a different division scheme from the ONIOM model real division.

As shown in eq 8, g^R is already including quadratic couplings with nonreaction-center variables implicitly. Hence, quadratic couplings are fully taken into account also in updated Hessian for reaction-center variables. It is known that usual Hessian update schemes by using g^R can be unstable when different coordinates are used for different parts,[29] e.g., redundant internal coordinates,[12−14] and Cartesian coordinates are used for reaction-center and nonreaction-center variables, respectively, in Gaussian programs. In the SHS technique, as discussed above, we use the angle coordinate for energy minimization on the scaled hyperspheres, and the same problem would happen if Hessian was updated in the angle coordinate. However, as explained above, we start with Cartesian Hessian (updated in Cartesian coordinate) in the process for determining minimization steps. Since microiterations are also performed in Cartesian coordinate in the Gaussian programs, we have never suffered from the instability problem in Hessian updates.

Only when exact Hessian is computed, a transformation of full-Hessian (including all of h^R, h^N, and h^C terms) into the effective Hessian of eq 9 is needed, where, in the present applications, exact Hessian was computed at the starting EQ point in the ADDF, every 50 macroiterations in both minimizations on scaled hyperspheres and macroiterations in EQ optimizations, every 5 macroiterations in TS optimizations, and every 10 macrosteps of IRC followings. In this transformation, we need the inverse matrix of Hessian for nonreaction-center variables. Here, we rewrite eq 9 as

$$\frac{\partial^2 V}{\partial x_i \partial x_j} = h_{ij}^R - \mathbf{h}_i^t \mathbf{H}_N^{-1} \mathbf{h}_j \quad (10)$$

where $\mathbf{h}_i = (h_{i1}^C, h_{i2}^C, ..., h_{iq}^C)$, and $\mathbf{H}_N$ is Hessian matrix for nonreaction-center variables. We have to perform either inversion of $\mathbf{H}_N$ or p time evaluations of $\mathbf{w}_j = \mathbf{H}_N^{-1} \mathbf{h}_j$. A
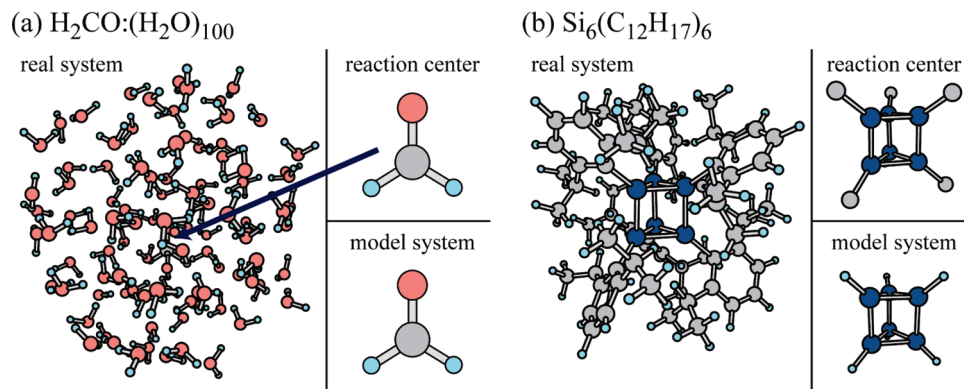
Figure 1. Definitions of real system, reaction-center, and ONIOM model system in the present applications to (a) $(H_2CO)(H_2O)_{100}$ and (b) $Si_6(C_{12}H_{17})_6$.

matrix inversion requires $O(q^3)$ costs, in general, whereas one $\mathbf{w}_j$ can be obtained by the iterative conjugate gradient method with (potentially) much less CPU time. Since p is much smaller than q, in general, and storage of p $\mathbf{w}_j$ vectors is trivial, the latter is used in this study. The most time-consuming part in the conjugate gradient method is the calculation of a product of $\mathbf{H}_N$ with an approximate vector for $\mathbf{w}_j$ in each iteration. Its cost can be reduced to $O(q)$ from $O(q^2)$ by a direct algorithm proposed recently.[29] This transformation requires $O(q^2)$ memory to keep $q \times q$ matrix elements of $\mathbf{H}_N$: ~0.45 GB for $M = 2\,500$ atoms ($q = 7\,500$), ~1.8 GB for $M = 5\,000$ atoms ($M = 15\,000$), ~7.2 GB for $M = 10\,000$ atoms ($q = 30\,000$). Hence, this transformation is trivial (relative to QM/MM Hessian calculations) in recent computers when nonreaction-center atoms are less than 5 000, although QM/MM Hessian calculation itself will be expensive in such large molecules. Memory size for this transformation also can be reduced significantly if the direct algorithm[29] is employed in the future. An $O(N)$ approach with iterative algorithm and inverse Hessian update for eigenvalues and eigenvectors search introduced in an earlier paper[85] will solve this problem as well.

III. Numerical Tests

In this paper, we show three numerical examples. The first two, for systems shown in Figure 1, are ONIOM applications in which the ONIOM model system is assumed to be the same as the μ-ADDF reaction center. One is reactions of formaldehyde surrounded by 100 water molecules with H_2CO as the ONIOM QM model system and the microiteration reaction center. Another is reactions of hexasilaprismane[86−88] in which Si_6C_6 are treated as reaction-center atoms in the microiteration, whereas Si_6H_6 with link H atoms is a ONIOM QM model system. The third example is a pure QM calculation for cyclohexanone (a cyclic ketone molecule with a C_6 ring) at the semiempirical PM3 level, where a carbonyl group and two carbons connected to the carbonyl group ($C_2C=O$) were treated as reaction-center atoms. Performance of the μ-ADDF calculations for the real systems was compared with performance of the ADDF calculations for the model systems.

$(H_2CO)(H_2O)_{100}$ in ONIOM. Total energy, gradient, and Hessian were computed by the ONIOM(HF/6-31G:AMBER)

Table 1. Harmonic Frequencies (in cm^{-1}) of the Effective Reaction-Center Hessian (H^{eff}) of $(H_2CO)(H_2O)_{100}$ for the Uncorrected Reaction-Center Hessian (h^R) of $(H_2CO)(H_2O)_{100}$ and for H_2CO

	effective Hessian, eq 10	uncorrected Hessian, h^R in eq 10	Hessian for H_2CO only
ω_1	1 332	1 348	1 329
ω_2	1 371	1 392	1 374
ω_3	1 675	1 690	1 674
ω_4	1 903	1 906	1 910
ω_5	3 170	3 226	3 208
ω_6	3 300	3 318	3 299

method by using the Gaussian program. Here, MM charges on H_2CO and H_2O are the natural charges of the isolated molecule and the original TIP3P parameters,[89] respectively, and they are mechanically embedded in the ONIOM calculations. The initial EQ structure was prepared as follows: (1) the center-of-mass position of H_2CO at HF/6-31G level was set to the origin, (2) one hundred water molecules with both random center-of-mass positions and orientations were distributed in a sphere centered at the origin, and (3) geometry optimization was performed from the structure of (2) by using an optimizer in the GRRM program coupled with the Gaussian microiteration code. Then, the μ-fADDF method was applied to this initial EQ structure. To see validity of the effective Hessian in eq 10, an μ-fADDF calculation with uncorrected Hessian (h^R in eq 10) was also performed for comparison. If the present μ-fADDF algorithm worked very well, then the numbers of macroiterations will be comparable to the numbers of total cycles in the fADD calculation for H_2CO, and the latter calculation at HF/6-31G level was also performed for comparison.

Table 1 compares harmonic frequencies of the effective Hessian, the uncorrected Hessian, and the model system. As seen in this table, frequencies of the effective Hessian are very similar to the ones for the model system, and differences in ω_1, ω_2, ω_3, and ω_6 are only 1−3 cm^{-1}. Although ω_4 (CO stretch) and ω_5 (symmetric CH stretch) show relatively large deviations, this may be because these two modes, especially ω_5, change the volume of H_2CO significantly, and such movements were suppressed by the surrounding waters. On the other hand, the uncorrected Hessian overestimates all the frequencies compared to those of the effective Hessian. This is because an energy-lowering effect via the microiteration

Transition Structure Explorer in Large Systems

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2739**

Table 2. Performance of the μ-fADDF for $(H_2CO)(H_2O)_{100}$ Compared to the μ-fADDF with the Uncorrected Hessian and the fADDF for the Model System

	μ-fADDF with effective Hessian, eq 10	μ-fADDF with uncorrected Hessian, h^R in eq 10	fADDF for H_2CO only
n_{ADD}	9	14	9
$n_{gradient}$	2 665	5 741	2 482
$n_{Hessian}$	117	192	101

(energy minimization for the nonreaction-center atoms) is omitted in the uncorrected Hessian, whereas it is correctly included in the effective Hessian via the quadratic couplings. How these frequency differences of $3-56$ cm^{-1} ($0.16-1.8\%$) affect the performance of the μ-fADDF will be discussed below.

Table 2 compares performance of the μ-fADDF with the effective Hessian, the μ-fADDF with the uncorrected Hessian, and the fADDF for the model system. Here, in all cases, a TS for both dissociation into $H_2 + CO$ and isomerization into hydroxymethylene were obtained. As expected, μ-fADDF with the effective Hessian showed almost the same performance as the model system. Exactly the same numbers of ADDs with the same characters were followed in these two cases: two isomerization pathways to hydroxymethylene via a TS, two direct dissociation routes to H + HCO, one direct dissociation route to O + CH$_2$, two dissociation routes to $H_2 + CO$ via a TS, and two dissociation routes to $H_2 + CO$ via a non-TS region with symmetric out-of-plane motion. The numbers of both QM gradient and Hessian calculations are also very similar to each other in these two cases. Although the μ-fADDF required bit more QM gradient and Hessian, this is because $(H_2CO)(H_2O)_{100}$ is not invariant against permutation of two H atoms in H_2CO. This symmetry breaking gives two independent TSs for migration of each H atom in the isomerization to hydroxymethylene and also two independent TSs for asymmetric association of two H atoms in the dissociation to $H_2 + CO$. Only two IRC calculations were required in the model system, whereas four calculations were required in the μ-fADDF case, since the IRC calculation is performed when an independent TS is found. This can be seen from the $n_{gradient}$ as well as $n_{Hessian}$ ratios between these two cases which are 1.07 and 1.16, respectively. As explained above, we perform Hessian calculations with a higher frequency in IRC calculations, and hence, the higher ratio in $n_{Hessian}$ implies that the difference was not caused in the ADDF process. It follows that the present μ-fADDF code coupled with a microiteration code in the Gaussian programs worked nearly perfectly in this very simple numerical test.

Only $0.16-1.8\%$ overestimation of the harmonic frequencies in the μ-fADDF with the uncorrected Hessian caused substantial deterioration in performance of the ADDF. Five extra pathways were followed via non-TS regions, and this increased the total numbers of gradient calculations to more than double. This is because minima on the scaled hyperspheres, based on wrong harmonic frequencies, are no longer ADD maximum points. This demonstrates how important the good harmonic reference is in the ADDF approach.

Another point we have to discuss here is that IRC calculations in the reverse direction from obtained TSs sometimes (three out of four TSs) did not come back to the initial EQ but lead to EQs for $(H_2CO)(H_2O)_{100}$ with other $(H_2O)_{100}$ arrangements. $(H_2CO)(H_2O)_{100}$ has numerous local structures with different $(H_2O)_{100}$ arrangements, and some of them are connected to one of the four TSs, while others are not. If one had to prepare correct $(H_2O)_{100}$ arrangements for each bond rearrangement, then the problem was much more complicated. However, fortunately, our observation in this test is that the μ-ADDF was able to find out all of the four TSs starting from an arbitrary $(H_2O)_{100}$ arrangement, which does not have direct connection to all four in full-dimension. This may be because small changes in the $(H_2O)_{100}$ arrangement caused only negligible effects on the reference harmonic frequencies of the reaction-center variables. No significant bumps were detected along traces of ADDs due to changes in $(H_2O)_{100}$ arrangements. A similar thing was also observed in the $Si_6(C_{12}H_{17})_6$ case. Hence, to our recommendation, one does not need to prepare a conformation of the nonreaction center parts with a direct connection to a final TS, but it is better to start from a conformation preferred at a given experimental condition or an experimental structure.

$Si_6(C_{12}H_{17})_6$ in ONIOM. Total energy, gradient, and Hessian were computed by the ONIOM(B3LYP/6-31G(d): UFF) method by using the Gaussian program. The initial EQ structure and MM-charges were prepared as follows: (1) a conformation sampling for the hydrocarbon arrangements with freezing the Si_6-cage structure was performed by using the lADDF method and the UFF force-field with zero MM-charges on the hydrocarbons, (2) MM charges were estimated by the QEq[90] method at the lowest energy conformer among about 100 ones obtained in (1) and mechanically embedded in the subsequent ONIOM calculations, and (3) geometry optimization at the ONIOM level was performed from the lowest energy conformer by using an optimizer in the GRRM program coupled with the Gaussian microiteration code. In the μ-lADDF calculation, to complete 15 of the largest ADDF, 45 ADDs were detected on the smallest hypersphere, and the 30 largest ones were followed among the 45. The 30 ADDs were followed simultaneously, and 15 are omitted without completion of the ADDF when the other 15 overcame barriers along their traces. The 15 largest ADDF were completed also in the lADDF calculation for the model system at B3LYP/6-31G(d) level. The μ-dADDF algorithm was also tested.

Figure 2a and b shows TSs and products for bond rearrangements in Si_6-cage backbone of hexasilaprizmane for both the real and the model system, respectively. Only the lowest energy TS is shown in Figure 2 when there are more than one TS for a bond rearrangement in the Si_6-cage due to different (a) hydrocarbon arrangement or (b) H atom direction. Although there are two other reaction pathways with H atom transfer in the model system, they are not shown in Figure 2b for simplicity. Among three TSs in Figure 2b, model-TS3 was taken from previous fADDF results[91] for explanation, although it was not found by the lADDF treating only 15 ADDs because of its high barrier. In Figure 2a, there
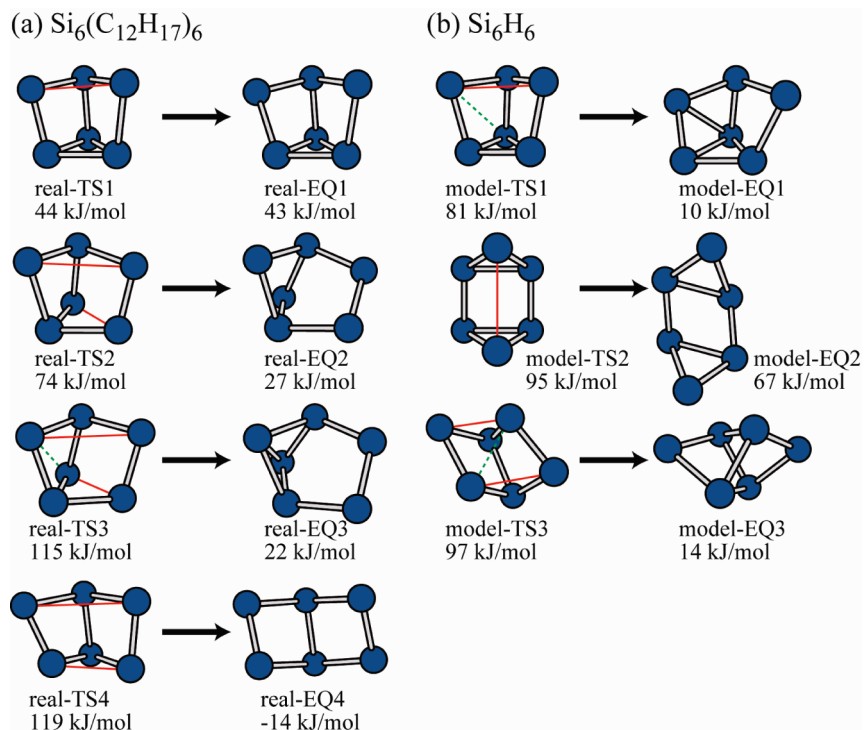
Figure 2. TSs and products of reactions in the Si_6 backbone for (a) $Si_6(C_{12}H_{17})_6$ and (b) Si_6H_6. Every Si atoms have one (a) $C_{12}H_{17}$ or (b) H. Thin (red) lines and dashed (green) lines on TSs show dissociating and generating bonds, respectively.

are four different Si−Si bond-breaking patterns in one of the triangles of the prism, depending on other subsequent bond breaking or formation: (1) the reaction stops after one bond breaking via real-TS1, (2) a backward Si−Si bond in another triangle breaks subsequently via real-TS2, (3) one bond formation accompanies in addition to the two bonds breaking like (2) via real-TS3, which leads to a product with hexasilabenzvalene backbone, (4) a Si−Si bond on the same side of another triangle breaks subsequently via real-TS4, which leads to a product with hexasila-Dewar-benzene backbone. The bond reorganization pattern in model-TS1 is similar to real-TS3, although another bond breaks in real-TS3 to reduce steric repulsion among hydrocarbon substituents in crowded parts. The bond reorganization pattern in model-TS3 is similar to the one in real-TS4, although further bond formation in model-TS3 is prohibited in the real-TS4 because of an increase of steric repulsion due to the bond formation. There are two other TSs in the model system involving H atom transfer from a Si to another Si, and their bond rearrangement pattern in the Si_6 backbone is very similar to both real-TS1 and -TS2, although $C_{12}H_{17}$ transfer in the real system is very difficult because of steric repulsion. A TS similar to model-TS2 was not found in the real system by the *l*ADDF treating of only 15 ADDs. Although it might be found also in the real system if *f*ADDF is performed, we did not do this test because of its high costs. This PES is related to the synthesis of $Si_6(C_{12}H_{17})_6$ with the hexasila-prizmane backbone and the thermal reaction from hexasila-Dewar-benzene to hexasilaprizmane after photolysis of hexasilaprizmane.[86−88] Such experimental studies aimed at synthesis of hexasilabenzene, which have not yet materialized,[92] and PESs between hexasilabenzene and other $Si_6(C_{12}H_{17})_6$ compounds are of great interest in inorganic

Table 3. Performance of μ-*l*ADDF for $Si_6(C_{12}H_{17})_6$, *l*ADDF for Si_6H_6 (model system), and μ-*d*ADDF for $Si_6(C_{12}H_{17})_6$

	μ-*l*ADDF	*l*ADDF	μ-*d*ADDF-TS (PR[a]-DB[b])	μ-*d*ADDF-TS (PR[a]-BV[c])	μ-*d*ADDF-EQ (DB[b]-BV[c])
n_{ADD}	15	15	1	1	1
$n_{gradient}$	14540	11020	146[d]/147[e]	82[d]/162[f]	189[e]/190[f]
$n_{Hessian}$	937	438	14[d]/16[e]	6[d]/8[f]	8[e]/8[f]

[a] Hexasilaprismane type EQ (see Figure 1). [b] Hexasila-Dewar-benzene type EQ (see real-eq 4 in Figure 2). [c] Hexasilabenzvalene type EQ (see real-eq 3 in Figure 2). [d] PR was the sphere-center. [e] DB was the sphere-center. [f] BV was the sphere-center.

chemistry. PESs for rearrangements of $Si_6(C_{12}H_{17})_6$ among hexasilaprizmane, hexasila-Dewar-benzene, hexasilabenzvalene, and hexasilabenzene at a higher computation level will be discussed in a forthcoming paper together with their theoretical photoabsorption spectra.

Table 3 shows performance of the μ-*l*ADDF and the *l*ADDF for the real system and the model system, respectively. Since we set the numbers of ADDs to be followed as 15 in the *l*ADDF treatment, n_{ADD} is 15 in both cases. There were many TSs for similar bond rearrangements in Si_6 due to C_1 symmetry in the real system, and this increased the numbers of IRC calculations significantly. Actually, 14 independent TSs were located in the real system, whereas only 5 TSs were located in the model system. In this system, we also tested the μ-*d*ADDF-TS and μ-*d*ADDF-EQ algorithms, which are applied to pathways among hexasilaprizmane, hexasila-Dewar-benzene, and hexasilabenzvalene. Table 3 also shows their performance, where costs for IRC calculations are not included in $n_{gradient}$ and $n_{Hessian}$ for μ-*d*ADDF. In the μ-*d*ADDF, n_{ADD} is always one. Although performance slightly depends on choices of sphere-center

in the μ-dADDF-TS between hexasilaprizmane and hexasila-Dewar-benzene and in the μ-dADDF-EQ between hexasila-Dewar-benzene and hexasilabenzvalene, very different in the μ-dADDF-TS between hexasilaprizmane and hexasilabenz-valene since hexasilabenzvalene is much closer to the TS than hexasilaprizmane. In the μ-dADDF-EQ, hexasilapriz-mane was obtained as unique intermediate, and the total cost for obtaining the pathway between hexasila-Dewar-benzene and hexasilabenzvalene by using the μ-dADDF methods is 417−499 gradient and 28−32 Hessian calculations.

$C_6H_{10}O$ in Pure QM Calculations. To see the perfor-mance of the μ-fADDF in the pure QM system, we applied it to cyclohexanone (a cyclic ketone molecule with a C_6 ring) at the semiempirical PM3 level, where a carbonyl group and two carbons connected to the carbonyl group ($C_2C=O$) were treated as reaction-center atoms. The μ-fADDF traced eight pathways: two isomerization pathways from $C_2C=O$ to C−C−O−C with a C_6O seven membered ring via a TS, two C−C bond dissociation pathways to a $(CH_2)_5CO$ diradical chain, a direct oxygen dissociation route, two dissociation routes to cyclopentane + CO via a TS, and a C−C bond dissociation route with a H atom transfer from CH_2 to $C=O$ generating $CH(CH_2)_4CHO$. Here, this result for C_2CO reaction-center is very similar to the H_2CO shown above. The number of macroiterations was 2 081 (1 996 force and 85 Hessian), and this is again comparable to the performance of the fADDF in H_2CO (see Table 2). In full-QM systems, costs of microiterations are not negligible, and 20 096 forces were required in total for microiteration. This implies that microiterations converged with only 10 iterations on average because of the following two reasons: (1) the maximum step size in each macroiteration was limited to 0.1 Å, and (2) the optimizer of Gaussian programs is very efficient. Hence, the numbers of force calculations in the present μ-ADDF applied to full-QM systems can be estimated to be ∼10 times larger than those of the model systems. This is a significant improvement over the fADDF since its cost will increase exponentially depending on the numbers of atoms.

IV. Remarks on Limitations and Future Applications

An upper limit in the numbers of nonreaction-center atoms, which can be handled in the present μ-ADDF algorithm (n_{MAX}), is determined by a memory size for Hessian storage in the transformation of nonreaction-center Hessian to the effective Hessian in eq 10. For example, n_{MAX} will be about 10 000 if available memory size is 8.0 GB. Hence, a small enzyme may be a target of future applications. Further increase of system size can be achieved by an introduction of an approximation in which a part of a huge molecule very far away from a reaction-center is frozen at an initially optimized or experimental geometry. Systematic search for TSs by the μ-ADDF with this approximation will give many candidates of TSs, and then they can be reoptimized in full dimension by using the microiteration-based optimizer in Gaussian programs. This treatment of dividing a huge molecule into three parts, i.e., reaction-center, nonreaction-center, and frozen atoms, is already available in the GRRM

program for future studies, although this approximation was not necessary in the present small systems.

When the numbers of nonreaction-center atoms are less than n_{MAX}, applicability is determined from the numbers of reaction-center QM atoms. Although it will strongly depend on available computer resources and computation levels, we expect from our previous applications that upper limits may be ∼20 atoms for fADDF, ∼50 atoms for lADDF, and >100 atoms in dADDF, for example, with DFT with a double-ζ basis set on eight or more cores of CPU. For example, in general, 100 or more atoms are often treated as a QM system in studies of enzyme reactions to take interactions between reaction-centers and surrounding residues and/or solvents into account. This is too much if a fully automated exploration of potential surface, by using the lADDF, is desired. In this case, use of the three-layer ONIOM(QM/QM/MM) with a semiempirical QM method for a medium layer could be used in an initial search for candidates of TS structures by the lADDF, where such a medium layer will also be treated by the microiteration. Then, they can be refined by using the two-layer ONIOM(QM/MM) and the standard TS optimizer. An application of the present approach to enzyme reactions will be a big challenge in the future.

Recently, we made systematic searches for minima on seams of crossing (MSX) structures possible for nonadiabatic transitions in photochemical and ion−molecule reactions by using the GRRM method.[93] A coupling of the MSX search code with the present μ-ADDF code will be an important development in the future.

V. Conclusion

In this study, we combined the microiteration technique with the GRRM method. Effective Hessian that includes full quadratic couplings between both reaction center and mi-croiteration variables were derived for the reaction center variables. The standard GRRM program was modified so that such an effective Hessian can be employed in the full and large-ADDF GRRM method as well as the double-ended GRRM method on the basis of μ-ADDF algorithm. Their performance was tested for ONIOM calculations for (1) H_2CO surrounded by 100 water molecules and (2) a prism-like Si_6-cage with six ($C_{12}H_{17}$) substituents, where H_2CO and Si_6C_6 were assumed to be reaction center atoms, respectively, as well as the pure QM calculations for cyclohexanone. The number of macroiterations in each application was compared to the number of total steps in similar calculations of the model systems, i.e., isolated H_2CO and Si_6H_6. The number of macroiterations was slightly larger than the number of total steps in the model system due to symmetry lowering (from C_{2v} to C_1) in H_2CO by the surrounding asymmetric $(H_2O)_{100}$ cluster. This effect was larger in the Si_6 case than in H_2CO because of a higher (D_{3h}) symmetry in the model system. Except for this, the μ-ADDF in both $(H_2CO)(H_2O)_{100}$ and $Si_6(C_{12}H_{17})$ worked almost as well as the conventional ADDF in the model systems in spite of flexible environ-ments. Therefore, we conclude that the GRRM method combined with the μ-ADDF algorithm will be a very powerful tool for exploring PESs of reaction-center variables in large molecular systems.

References

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

(2) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; Wiley: Chichester, U.K., 2007.

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

(4) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.

(5) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.

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

(7) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461–462*, 1.

(8) Morokuma, K. *Bull. Korean Chem. Soc.* **2003**, *24*, 797.

(9) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214.

(10) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800.

(11) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52.

(12) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. *J. Am. Chem. Soc.* **1979**, *101*, 2550.

(13) Fogarasi, G.; Zhou, X.; Taylor, P.; Pulay, P. *J. Am. Chem. Soc.* **1992**, *114*, 8191.

(14) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49.

(15) (a) Broyden, C. G. *J. Inst. Math. Appl.* **1970**, *6*, 76. (b) Fletcher, R. *Computer Journal (Switzerland)* **1970**, *13*, 317. (c) Goldfarb, D. *Mathematics of Computation* **1970**, *24*, 23. (d) Shanno, D. F. *Mathematics of Computation* **1970**, *24*, 647.

(16) Murtagh, B.; Sargent, R. W. H. *Computer Journal (Switzerland)* **1972**, *13*, 185.

(17) Powell, M. J. D. *Mathematical Programming* **1971**, *1*, 26.

(18) Bofill, J. M. *J. Comput. Chem.* **1994**, *15*, 1.

(19) Golab, J. T.; Yeager, D. L.; Jørgensen, P. *Chem. Phys.* **1983**, *78*, 175.

(20) Helgaker, T. *Chem. Phys. Lett.* **1991**, *182*, 503.

(21) Culot, P.; Dive, G.; Nguyen, V. H.; Ghuysen, J. M. *Theor. Chim. Acta.* **1992**, *82*, 189.

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

(23) Zhang, Y. K.; Liu, H. Y.; Yang, W. T. *J. Chem. Phys.* **2000**, *112*, 3483.

(24) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442.

(25) Hayashi, S.; Ohmine, I. *J. Phys. Chem.* **2000**, *104*, 10678.

(26) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.

(27) Hall, R. J.; Hindle, S. A.; Burton, N. A.; Hillier, I. H. *J. Comput. Chem.* **2000**, *21*, 1433.

(28) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760.

(29) Vreven, T.; Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701.

(30) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87.

(31) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325.

(32) Kudin, K. N.; Scuseria, G. E. *Chem. Phys. Lett.* **1998**, *283*, 61.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(34) Lundberg, M.; Kawatsu, T.; Vreven, T.; Frisch, M.; Morokuma, K. *J. Chem. Theory Comput.* **2009**, *5*, 222.

(35) Müller, K.; Brown, L. D. *Theor. Chim. Acta* **1979**, *53*, 75.

(36) Peng, C.; Schlegel, H. B. *Isr. J. Chem.* **1993**, *33*, 449.

(37) Ayala, P. Y.; Schlegel, H. B. *J. Chem. Phys.* **1997**, *107*, 375.

(38) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978.

(39) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B: Condens. Matter* **2002**, *66*, 052301.

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

(41) Quapp, W. *J. Chem. Phys.* **2005**, *122*, 174106.

(42) Maeda, S.; Ohno, K. *Chem. Phys. Lett.* **2005**, *404*, 95.

(43) Maeda, S.; Ohno, K. *J. Chem. Phys.* **2006**, *124*, 174306.

(44) Pancíř, J. *Collect. Czech. Chem. Commun.* **1975**, *40*, 1112.

(45) Basilevsky, M. V.; Shamov, A. G. *Chem. Phys.* **1981**, *60*, 347.

(46) Basilevsky, M. V. *Chem. Phys.* **1982**, *67*, 337.

(47) Rowe, D. J.; Ryman, A. *J. Math. Phys.* **1982**, *23*, 732.

(48) Hoffman, D. K.; Nord, R. S.; Ruedenberg, K. *Theor. Chim. Acta* **1986**, *69*, 265.

(49) Jørgensen, P.; Jensen, H. J. A.; Helgaker, T. *Theor. Chim. Acta* **1988**, *73*, 55.

(50) Quapp, W. *Theor. Chim. Acta* **1989**, *75*, 447.

(51) Schlegel, H. B. *Theor. Chim. Acta* **1992**, *83*, 15.

Transition Structure Explorer in Large Systems

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2743**

(52) Sun, J.-Q.; Ruedenberg, K. *J. Chem. Phys.* **1993**, *98*, 9707.

(53) Bondensgård, K.; Jensen, F. *J. Chem. Phys.* **1996**, *104*, 8025.

(54) Quapp, W.; Hirsch, M.; Imig, O.; Heidrich, D. *J. Comput. Chem.* **1998**, *19*, 1087.

(55) Quapp, W.; Hirsch, M.; Heidrich, D. *Theor. Chem. Acc.* **1998**, *100*, 285.

(56) Bofill, J. M.; Anglada, J. M. *Theor. Chem. Acc.* **2001**, *105*, 463.

(57) Crehuet, R.; Bofill, J. M.; Anglada, J. M. *Theor. Chem. Acc.* **2002**, *107*, 130.

(58) Hirsch, M.; Quapp, W. *J. Comput. Chem.* **2002**, *23*, 887.

(59) Dallos, M.; Lischka, H.; Monte, E. V. D.; Hirsch, M.; Quapp, W. *J. Comput. Chem.* **2002**, *23*, 576.

(60) Hirsch, M.; Quapp, W. *J. Mol. Struct. (Theochem)* **2004**, *689*, 1.

(61) Ohno, K.; Maeda, S. *Chem. Phys. Lett.* **2004**, *384*, 277.

(62) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2005**, *109*, 5742.

(63) Ohno, K.; Maeda, S. *J. Phys. Chem. A* **2006**, *110*, 8933.

(64) Ohno, K.; Maeda, S. *Phys. Scr.* **2008**, *78*, 058122.

(65) Yang, X.; Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2005**, *109*, 7319.

(66) Yang, X.; Maeda, S.; Ohno, K. *Chem. Phys. Lett.* **2006**, *418*, 208.

(67) Yang, X.; Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 5099.

(68) Watanabe, Y.; Maeda, S.; Ohno, K. *Chem. Phys. Lett.* **2007**, *447*, 21.

(69) Maeda, S.; Ohno, K. *Chem. Phys. Lett.* **2008**, *460*, 55.

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

(71) Ishida, K.; Morokuma, K.; Komornicki, A. *J. Chem. Phys.* **1977**, *66*, 2153.

(72) Page, M.; McIver, J. W., Jr. *J. Chem. Phys.* **1988**, *88*, 922.

(73) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.

(74) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 4527.

(75) Bell, R. P. *Proc. R. Soc. London, Ser. A* **1936**, *154*, 414.

(76) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1936**, *32*, 1333.

(77) Luo, Y.; Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 10732.

(78) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2008**, *112*, 2962.

(79) Luo, Y.; Maeda, S.; Ohno, K. *J. Comput. Chem.* **2009**, *30*, 952.

(80) Maeda, S.; Ohno, K. *J. Phys. Chem. A* **2007**, *111*, 13168.

(81) Maeda, S.; Ohno, K. *J. Am. Chem. Soc.* **2008**, *130*, 17228.

(82) Maeda, S.; Ohno, K. *Chem. Phys. Lett.* **2003**, *381*, 177.

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

(84) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.

(85) Farkas, Ö.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 10806.

(86) Sekiguchi, A.; Yatabe, T.; Kabuto, C.; Sakurai, H. *J. Am. Chem. Soc.* **1994**, *115*, 5853.

(87) Sekiguchi, A.; Yatabe, T.; Doi, S.; Sakurai, H. *Phosphorus, Sulfur Silicon Relat. Elem.* **1994**, *93−94*, 193.

(88) Sekiguchi, A.; Sakurai, H. *Adv. Organomet. Chem.* **1995**, *37*, 1.

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

(90) Rappé, A. K.; Goddard III, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.

(91) Moteki, M.; Maeda, S.; Ohno, K. *Organometallics* **2009**, *28*, 2218.

(92) Lee, V. Y.; Sekiguchi, A. *Chem. Soc. Rev.* **2008**, *37*, 1652.

(93) Maeda, S.; Ohno, K.; Morokuma, K. *J. Phys. Chem. A* **2009**, *113*, 1704.

JCTC Journal of Chemical Theory and Computation

Comparison of Global Reactivity Descriptors Calculated Using Various Density Functionals: A QSAR Perspective

R. Vijayaraj,[†] V. Subramanian,*,[†] and P. K. Chattaraj*,[‡]

Chemical Laboratory, Central Leather Research Institute, Council of Scientific and Industrial Research, Adyar, Chennai 600 020, India, and Department of Chemistry, Indian Institute of Technology, Kharagpur 721302, India

Abstract: Conceptual density functional theory (DFT) based global reactivity descriptors are used to understand the relationship between structure, stability, and global chemical reactivity. Furthermore, these descriptors are employed in the development of quantitative structure−activity (QSAR), structure−property (QSPR), and structure−toxicity (QSTR) relationships. However, the predictive power of various relationships depends on the reliable estimates of these descriptors. The basic working equations used to calculate these descriptors contain both the ionization potential and the electron affinity of chosen molecules. Therefore, efficiency of different density functionals (DFs) in predicting the ionization potential and the electron affinity has to be systematically evaluated. With a view to benchmark the method of calculation of global reactivity descriptors, comprehensive calculations have been carried out on a series of chlorinated benzenes using a variety of density functionals employing different basis sets. In addition, to assess the utility of global reactivity descriptors, the relationships between the reactivity−electrophilicity and the structure−toxicity have been developed. The ionization potential and the electron affinity values obtained from M05-2X method using the ΔSCF approach are closer to the corresponding experimental values. This method reliably predicts these electronic properties when compared to the other DFT methods. The analysis of a series of QSTR equations reveals that computationally economic DFT functionals can be effectively and routinely applied in the development of QSAR/QSPR/QSTR.

Introduction

The conceptual DFT based global reactivity and the local reactivity descriptors are useful quantities to understand the global reactivity and the local site selectivity, respectively. Reliable calculation of global reactivity descriptors is important in conceptual DFT, as these values are used as input for understanding the relationships between structure, stability, and reactivity as well as for QSAR/QSPR/QSTR development.[1,2] The working equations used to calculate the various global reactivity descriptors involve the ionization potential and the electron affinity of selected molecules.

Computationally, the ionization potential and the electron affinity are calculated from total electronic energy calculations on the $N - 1$, N, and $N + 1$ electron systems at the neutral geometry. This method is referred to as "delta SCF" or "ΔSCF" method. Since, this approach is computationally demanding, these quantities are also calculated using Koopmans' approximation.[3] However, this approximation is suitable for molecules with positive electron affinity but is less appropriate for systems with negative electron affinity. De Proft et al. have addressed this issue, and various new working relations have been proposed to compute absolute hardness from the ionization potential and the electron affinity.[4,5]

Both conventional ab initio wave function theory[6] and Kohn−Sham DFT methods[7] have been employed to predict the ionization potential and the electron affinity and, thus,

* Corresponding authors. E-mail: subuchem@hotmail.com (V.S.) and pkc@chem.iitkgp.ernet.in (P.K.C.).

† Central Leather Research Institute.

‡ Indian Institute of Technology.

Table 1. List of Different DFT Methods Evaluated in This Study

functional	type	percentage of Hartree−Fock exchange	exchange/correlation functional	ref
SVWN5	LSDA	0	Slater's local ex/VWN no.5	41−43
BLYP	GGA	0	Becke88/Lee−Yang−Parr	44, 45
B3LYP	HGGA	20	Becke88/Lee−Yang−Parr	44−46
MPWB1K	HMGGA	44	modified PW91/Becke95	15, 47, 48
BB1K	HMGGA	42	Becke88/Becke95	16, 44, 47
MPW1B95	HMGGA	31	modified PW91/Becke95	15, 47, 48
MO5	HMGGA	28	self-correlation-free	17
MO5-2X	HMGGA	56	self-correlation-free	18

global reactivity descriptors. Prediction of these electronic properties using wave function theory and DFT approaches has several limitations.[8,9] In this context, DFT methods have been shown to be more efficient than conventional ab initio methods due to their excellent cost-to-performance ratio.[10,11] The most popular hybrid DFT based B3LYP methods have been employed in several studies to predict the global and local reactivity descriptors.[12]

Numerous studies have been carried out to develop new exchange and correlation functionals to enhance the quality of prediction and expand the applicability of DFT methods.[7,13] Scuseria and Staroverov have reviewed the various strategies used to refine functionals for DFT.[14] Truhlar and co-workers have made seminal contributions to the development and refinement of functionals with wide applicability in chemistry using diverse set of training and test sets.[8,10,11,13,15−18] In general, two different approaches have been used for the development of new functionals, viz, nonempirical, and semiempirical. A nonempirical approach leads to the development of ab initio DFT (or pure DFT) in which functionals are constructed from the first principles with exact constraints. In the semiempirical DFT approach, the new functionals depend on the one or more parameters which are obtained from fitting to different sets of experimental data. The following strategies have been applied to design and develop different new density functionals: (i) local spin density approximation (LSDA), (ii) generalized gradient approximation (GGA), (iii) constraint satisfaction, (iv) modeling exchange−correlation hole, (v) empirical fits, and (vi) mixing Hartree−Fock and approximate DFT exchanges.

Different functionals used in the literature are referred to as various rungs of "Jacob's ladder".[19,20] They are classified on the basis of the number and the types of the constituents used in the functionals. The lowest rung is the LSDA in which the density functional depends only on density, and the second rung is the GGA in which the density functional depends on density and its reduced gradient. The third rung is meta GGA in which the functional also depends on kinetic energy density. The fourth rung is hybrid GGA, which employs some percentage of HF exchange. There are two kinds of DFT methods that fit into the fourth rung of the Jacob's ladder, and they are called hybrid GGA (a combination of GGA with Hartree−Fock exchange). The typical examples are B3LYP, PBE0, and MPW1K. The other type is hybrid meta GGA (a combination of meta GGA with Hartree−Fock exchange). The MPWB1K, MPW1B95, and TPSSh functionals belong to this category. Both hybrid GGA and hybrid meta GGA are semiempirical, and they have been

very successful in chemistry. Recently, new density functionals have been developed by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions.[18] These functionals are denoted as M05 and M05-2X. Subsequently, the same authors have developed M06, M06-2X, M06-HF, and M06-2L using large energetic databases. Since several functionals have been developed in the last five years, it is necessary to analyze the performance of various categories of DFT methods to predict the global reactivity descriptors. In this context, the DFT method which reliably predicts the ionization potential and the electron affinity is necessary. Hence, we have selected functionals from different categories. The selected functionals are presented in Table 1. Various basis sets such as 6-31G*, 6-31G**, 6-31+G*, 6-31+G**, 6-31++G**, 6-311++G**, aug-CC-PVDZ, and aug-CC-PVTZ have been chosen in this study. To explore the performance of various new functionals, the calculated ionization potential and electron affinity values have been compared with available experimental values. To benchmark the appropriate method for the calculation of global reactivity descriptors from QSAR/QSPR/QSTR perspectives, a series of chlorine-substituted benzenes have been considered as model systems, and QSTR has been developed for fish toxicities of these compounds against *Poecilia reticulate*.[21]

Theoretical Background. The chemical potential (μ) and chemical hardness (η) are defined within DFT as the first- and second-order derivatives, respectively, of the Kohn−Sham (KS) energy to the number of particles at constant potential $v(\bar{r})$.[3,22−24]

$$\mu = -\chi = (\partial E/\partial N)_{v(\bar{r})} \qquad (1)$$

$$\eta = (\partial^2 E/\partial N^2)_{v(\bar{r})} = (\partial \mu/\partial N)_{v(\bar{r})} \qquad (2)$$

where E and N are the total energy and the number of electrons in the system, respectively, and $v(\bar{r})$ is the external potential. The global hardness is an indicator of the overall stability of the system. A maximum hardness principle (MHP) relating hardness to stability at a constant chemical potential has been proposed by Pearson.[3,22,25] The proof of the same has been provided by Parr and Chattaraj.[26−29] The chemical potential, μ, is identified as the negative of the electronegativity (χ) by Iczkowski and Margrave.[30] Although, these quantities have been defined based on the KS energy, the same definitions have been routinely used in the canonical molecular orbital theory. The most popularly used formulas for the computation of these quantities use the three point finite difference approximation

and express μ and η through the ionization potential (I) and the electron affinity (A).

$$\mu = -\chi = -\frac{1}{2}(I + A) \tag{3}$$

$$\eta = (I - A) \tag{4}$$

Ionization potential and electron affinity can be calculated using ΔSCF method using the formulas given below:

$$I \approx E(N - 1) - E(N) \tag{5}$$

$$A \approx E(N) - E(N + 1) \tag{6}$$

where I and A are obtained from total electronic energy calculations on the $N - 1$, N, and $N + 1$ electron systems at the neutral geometry. Another DFT approximation originating from the Koopmans' theorem[3] is

$$I \approx -\varepsilon_{HOMO}, A \approx -\varepsilon_{LUMO} \tag{7}$$

where ε_{HOMO} and ε_{LUMO} are the KS one electron eigenvalues associated with the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), respectively, from approximate DFT calculations on neutral molecules. These approximations have also been widely used in the canonical molecular orbital theory. Various problems in the calculation of μ and η have been highlighted in a recent study.[4] Other methods of calculation of global hardness have been reported in the literature.[4,31−33] Compact finite difference schemes with spectral-like resolution have been used to calculate hardness as second derivatives of energy with respect to the electron number.[31] An algorithm for computation of density functional based orbital reactivity indices, such as orbitally resolved hardness and softness tensors, and total hardness, and the Fukui functions have been extended to degenerate electronic states.[32] Recently, Tozer and De Proft have described the problems associated with the calculation of absolute hardness.[4] A simple correction to the Koopmans' expression is highlighted on the basis of a consideration of integer discontinuity. The resulting expression does not require the explicit computation of the electron affinity. The alternative expression for the hardness provided by Tozer and De Proft is

$$\eta = \frac{\varepsilon_{LUMO} - \varepsilon_{HOMO}}{2} + \varepsilon_{HOMO} + I \tag{8}$$

where $I = I_0$ is the experimental ionization potential. It can be further written as

$$\eta = \frac{\varepsilon_{LUMO} + \varepsilon_{HOMO}}{2} + I \tag{9}$$

Although there has not been a rigorous definition, the concept of electrophilicity has been known for several decades. In 1999, Parr et al. proposed a rigorous definition based on the energy lowering associated with a maximum amount of electron flow between two species.[34] It is defined as

$$\omega = \frac{\mu^2}{2\eta} \tag{10}$$
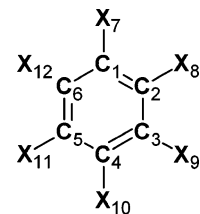
Figure 1. Chlorobenzene template structure with atom numbering.

Subsequently, a number of studies have been made to understand the global reactivity and site selectivity of systems using this new index. A detailed account on various aspects of this new electrohilicity index has been summarized in a recent review and update.[35]

Computational Details. The geometries of the benzene and 12 chlorine-substituted benzenes are optimized using various DFT methods listed in Table 1 with 6-31G*, 6-31G**, 6-31+G*, 6-31+G**, 6-31++G**, 6-311++G**, aug-CC-PVDZ, and aug-CC-PVTZ basis sets. The harmonic vibrational frequencies are calculated for all the geometries, which confirm that the obtained structures are minima on the potential energy surface. The global electronic properties are calculated using the eqs 3, 4, and 10. Both the ionization potential and electron affinity values are calculated using ΔSCF procedure (eqs 5 and 6) and Koopmans' theorem (eq 7). With a view to assess how various reactivity indices evolve, the global reactivity descriptors computed from different schemes are also compared with Tozer and De Proft's expression for η (eq 8). All calculations are carried out using Gaussian 03 suite of programs.[36]

Results and Discussion

Various chlorinated benzenes (CBs) selected in this study are schematically represented in Figure 1 along with the atom numbering. Since, the objective of the present investigation is to explore efficiencies of various DFs, a systematic analysis has been carried out on the structure, the energy, and the electronic properties of chlorobenzenes.

Energetics and Structural Details of Chlorobenzenes. It is well-known that the sterical interaction between any two adjacent chlorine atoms in substituted benzenes decreases the stability of the respective isomers. As a consequence, the lowest stability can be expected for the CBs with chlorine at ortho and meta positions. The calculated total energies, the thermodynamic quantities (enthalpy and free energy), and the relative energies using various methods with 6-31++G** basis set for all chlorobenzenes along with benzene are presented in Tables 2−4. The stability of a conformer with a given number of chlorine atoms is determined by its relative energy (ΔE) with respect to the most stable isomer. For dichlorobenzenes, all the chosen methods predict the para isomer (1,4-C2B) of dichlorobenzene as the most stable isomer, except in the SVWN5 method. It can be found that SVWN5 method predicts 1,3-C2B as the most stable isomer (Table 2). This discrepancy may be attributed to the problems in the treatment of exchange and correlation interactions

A QSAR Perspective

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2747**

Table 2. Calculated Energies E (Hartree) and Relative Energy (ΔE, kcal/mol) with respect to the Most Stable Isomer with a Given Number of Chlorine Atoms using 6-31++G** Basis Set

system	B3LYP E	B3LYP ΔE	MPW1B95 E	MPW1B95 ΔE	MPWB1K E	MPWB1K ΔE	BB1K E	BB1K ΔE	SVWN5 E	SVWN5 ΔE	BLYP E	BLYP ΔE	M05 E	M05 ΔE	M05-2X E	M05-2X ΔE
CB	−691.86		−691.82		−691.82		−691.81		−689.37		−691.74		−691.68		−691.81	
1,2-C2B	−1151.45	2.44	−1151.47	1.85	−1151.47	1.85	−1151.45	1.95	−1147.82	1.31	−1151.32	2.47	−1151.28	2.24	−1151.38	1.94
1,3-C2B	−1151.46	0.05	−1151.47	0.01	−1151.47	0.00	−1151.46	0.01	−1147.83	0.01	−1151.32	0.01	−1151.28	0.06	−1151.38	0.08
1,4-C2B	−1151.46	0.00	−1151.47	0.00	−1151.47	0.00	−1151.46	0.00	−1147.83	0.02	−1151.32	0.00	−1151.28	0.00	−1151.38	0.00
1,2,3-C3B	−1611.04	4.74	−1611.12	3.74	−1611.12	3.64	−1611.09	3.96	−1606.27	2.44	−1610.89	4.86	−1610.88	4.21	−1610.95	3.65
1,2,4-C3B	−1611.04	2.14	−1611.12	1.73	−1611.12	1.64	−1611.09	1.83	−1606.27	1.13	−1610.90	2.22	−1610.88	1.89	−1610.95	1.63
1,3,5-C3B	−1611.05	0.00	−1611.12	0.00	−1611.12	0.00	−1611.10	0.00	−1606.28	0.00	−1610.90	0.00	−1610.88	0.00	−1610.95	0.00
1,2,3,4-C4B	−2070.63	2.82	−2070.76	2.13	−2070.76	2.15	−2070.73	2.27	−2064.72	1.33	−2070.47	2.86	−2070.48	2.52	−2070.51	2.10
1,2,3,5-C4B	−2070.63	0.40	−2070.77	0.33	−2070.77	0.34	−2070.73	0.36	−2064.72	0.19	−2070.47	0.36	−2070.48	0.30	−2070.52	0.30
1,2,4,5-C4B	−2070.63	0.00	−2070.77	0.00	−2070.77	0.00	−2070.73	0.00	−2064.72	0.00	−2070.48	0.00	−2070.48	0.00	−2070.52	0.00
1,2,3,4,5-C5B	−2530.21		−2530.41		−2530.41		−2530.37		−2523.17		−2530.05		−2530.07		−2530.08	
1,2,3,4,5,6-C6B	−2989.79		−2990.05		−2990.05		−2990.00		−2981.61		−2989.62		−2989.67		−2989.64	
benzene	−232.27		−232.17		−232.16		−232.17		−230.92		−232.15		−232.07		−232.23	

Table 3. Enthalpies H (Hartree) and Free Energies G (Hartree) Calculated from B3LYP, MPW1B95, MPWB1K, and BB1K Methods with 6-31++G** Basis Set[a]

system	B3LYP H	B3LYP ΔH	B3LYP G	B3LYP ΔG	MPW1B95 H	MPW1B95 ΔH	MPW1B95 G	MPW1B95 ΔG	MPWB1K H	MPWB1K ΔH	MPWB1K G	MPWB1K ΔG	BB1K H	BB1K ΔH	BB1K G	BB1K ΔG
CB	−691.77		−691.80		−691.72		−691.76		−691.72		−691.76		−691.72		−691.75	
1,2-C2B	−1151.36	2.45	−1151.40	2.55	−1151.38	1.91	−1151.42	1.96	−1151.38	1.91	−1151.42	1.80	−1151.36	2.00	−1151.41	1.93
1,3-C2B	−1151.37	0.04	−1151.41	0.04	−1151.38	0.00	−1151.42	0.00	−1151.38	0.00	−1151.42	0.00	−1151.37	0.00	−1151.41	0.00
1,4-C2B	−1151.37	0.00	−1151.41	0.00	−1151.38	0.02	−1151.42	0.09	−1151.38	0.03	−1151.42	0.15	−1151.37	0.02	−1151.41	0.13
1,2,3-C3B	−1610.96	4.82	−1611.00	5.03	−1611.04	3.80	−1611.08	4.07	−1611.04	3.79	−1611.08	4.05	−1611.01	4.03	−1611.05	4.27
1,2,4-C3B	−1610.96	2.19	−1611.01	2.28	−1611.04	1.73	−1611.08	1.76	−1611.04	1.72	−1611.08	1.60	−1611.01	1.84	−1611.06	1.73
1,3,5-C3B	−1610.97	0.00	−1611.01	0.00	−1611.04	0.00	−1611.09	0.00	−1611.04	0.00	−1611.09	0.00	−1611.02	0.00	−1611.06	0.00
1,2,3,4-C4B	−2070.55	2.75	−2070.60	2.87	−2070.69	2.08	−2070.74	2.13	−2070.69	2.10	−2070.74	2.09	−2070.66	2.22	−2070.70	2.22
1,2,3,5-C4B	−2070.56	0.37	−2070.60	0.41	−2070.70	0.32	−2070.74	0.34	−2070.69	0.34	−2070.74	0.36	−2070.66	0.35	−2070.71	0.38
1,2,4,5-C4B	−2070.56	0.00	−2070.61	0.00	−2070.70	0.00	−2070.74	0.00	−2070.69	0.00	−2070.74	0.00	−2070.66	0.00	−2070.71	0.00
1,2,3,4,5-C5B	−2530.15		−2530.20		−2530.35		−2530.40		−2530.35		−2530.40		−2530.30		−2530.35	
1,2,3,4,5,6-C6B	−2989.74		−2989.79		−2990.00		−2990.05		−2990.00		−2990.05		−2989.94		−2990.00	
benzene	−232.16		−232.20		−232.06		−232.09		−232.06		−232.09		−232.06		−232.10	

[a] The relative energies of ΔH and ΔG (in kcal/mol) are with respect to the stable isomer of given number of chlorine atoms.

when compared to other functionals. In fact, it can be noted that the Hartree−Fock exchange in SVWN5 and BLYP is zero. However, BLYP predicts the trend which is similar to that of other functionals and is discussed in the subsequent section. Hence, the inherent features of Slater's exchange and the VWN's correlation functionals may be responsible for the incorrect experimental trend. Further, the results clearly show that an intricate balance between the exchange and correlation terms is necessary for the appropriate prediction of the structure.

The calculated energy difference between the para and the ortho isomers from MPW1B95, MPWB1K, BB1K, and M05-2X methods are similar. It ranges from 1.85 to 1.95 kcal/mol. When compared to these methods, B3LYP, BLYP, and M05 methods predict marginally higher energy differences of 2.24 to 2.47 kcal/mol. Similarly, 1,3,5-C3B and 1,2,4,5-C4B are shown to be the most stable isomers, whereas 1,2,3-C3B and 1,2,3,4-C4B are the least stable isomers for the tri- and tetrachlorobenzenes, respectively. The predicted energy difference between the isomers of tri- and tetrachlorobenzenes follows the same trend as that observed for dichlorobenzenes. It has been found from the above analysis that the MPW1B95, MPWB1K, BB1K, and M05-2X methods predict similar trends, whereas B3LYP, BLYP, and M05 methods show similar trends, which is different from the former. The SVWN5 predicted energy does not show any agreement with the other chosen methods. The

thermodynamic quantities, such as enthalpy and free energy, as well as their relative values are shown in Tables 3−4. The trends observed in the ΔH and ΔG calculations are more or less parallel to that of ΔE.

Tables S1−S2 shown in the Supporting Information provides the details of both the calculated bond lengths and angles using various methods with 6-31++G** basis set along with the experimental values for comparison.[37,38] It is evident from the tables that the calculated geometrical properties are in good agreement with the experimental values. It can be found from the Supporting Information (Table S1) that chlorine substitution in the C_1 reduces the bond length of the C_1-C_2 and C_1-C_6 to 0.002 Å, which is observed in all the methods. The bond lengths predicted by the M05 and M05-2X methods are closer to the experimental values when compared to the other methods. The BLYP method slightly overestimates the bond distances.

Ionization Potential and Electron Affinity. The ionization potentials and the electron affinities calculated from various methods using 6-31++G** basis set employing Koopmans' theorem are shown in Table 5, and the results obtained from the remaining basis functions are shown in Supporting Information (Tables S3−S8). The experimental ionization potentials and the electron affinities of CBs obtained from the NIST standard reference database are also presented in the same table. It can be found that all selected

Table 4. Enthalpies *H* (Hartree) and Free Energies *G* (Hartree) Calculated from SVWN5, BLYP, M05, and M05-2X Methods with 6-31++G** Basis Set[a]

| | SVWN5 | | | | BLYP | | | | M05 | | | | M05-2X | | | |
system	*H*	Δ*H*	*G*	Δ*G*	*H*	Δ*H*	*G*	Δ*G*	*H*	Δ*H*	*G*	Δ*G*	*H*	Δ*H*	*G*	Δ*G*
CB	−689.28		−689.31		−691.64		−691.68		−691.58		−691.62		−691.71		−691.75	
1,2-C2B	−1147.74	1.39	−1147.78	1.51	−1151.23	2.50	−1151.27	2.60	−1151.19	2.29	−1151.23	2.45	−1151.29	1.96	−1151.33	2.03
1,3-C2B	−1147.74	0.00	−1147.78	0.00	−1151.23	0.00	−1151.27	0.00	−1151.19	0.08	−1151.23	0.16	−1151.29	0.01	−1151.33	0.00
1,4-C2B	−1147.74	0.02	−1147.78	0.02	−1151.23	0.01	−1151.27	0.01	−1151.19	0.00	−1151.23	0.00	−1151.29	0.00	−1151.33	0.01
1,2,3-C3B	−1606.19	2.64	−1606.24	2.88	−1610.82	4.95	−1610.86	5.17	−1610.80	4.35	−1610.84	4.67	−1610.87	3.69	−1610.91	3.75
1,2,4-C3B	−1606.19	1.22	−1606.24	1.32	−1610.82	2.27	−1610.86	2.36	−1610.80	1.97	−1610.84	2.12	−1610.87	1.67	−1610.91	1.70
1,3,5-C3B	−1606.20	0.00	−1606.24	0.00	−1610.82	0.00	−1610.87	0.00	−1610.80	0.00	−1610.85	0.00	−1610.87	0.00	−1610.91	0.00
1,2,3,4-C4B	−2064.65	1.30	−2064.69	1.42	−2070.40	2.78	−2070.45	2.89	−2070.40	2.51	−2070.45	2.69	−2070.44	2.08	−2070.49	2.30
1,2,3,5-C4B	−2064.65	0.18	−2064.70	0.22	−2070.40	0.34	−2070.45	0.37	−2070.41	0.34	−2070.45	0.50	−2070.44	0.25	−2070.49	0.39
1,2,4,5-C4B	−2064.65	0.00	−2064.70	0.00	−2070.41	0.00	−2070.45	0.00	−2070.41	0.00	−2070.45	0.00	−2070.44	0.00	−2070.49	0.00
1,2,3,4,5-C5B	−2523.10		−2523.15		−2529.98		−2530.03		−2530.01		−2530.06		−2530.02		−2530.07	
1,2,3,4,5,6-C6B	−2981.56		−2981.61		−2989.56		−2989.62		−2989.61		−2989.66		−2989.59		−2989.64	
benzene	−230.82		−230.85		−232.05		−232.08		−231.97		−232.00		−232.13		−232.16	

[a] The relative energies of Δ*H* and Δ*G* (in kcal/mol) are with respect to the stable isomer of given number of chlorine atoms.

Table 5. Ionization Potential (*I*) and Electron Affinity (*A*) of Chlorinated Benzenes in eV Calculated from Various Methods with 6-31++G** Basis Set using Koopmans' Theorem

| | expt[a] | | B3LYP | | MPWB1K | | MPW1B95 | | BB1K | | SVWM | | BLYP | | M05 | | M05-2X | |
system	*I*	*A*	*I*	*A*	*I*	*A*	*I*	*A*	*I*	*A*	*I*	*A*	*I*	*A*	*I*	*A*	*I*	*A*
CB	9.07 ± 0.02	−[c]	6.95	0.80	6.02	1.40	6.02	1.43	5.98	1.35	6.76	2.13	5.97	1.41	7.15	0.49	8.34	−0.32
1,2-C2B	9.06 ± 0.02	0.093	7.07	1.08	6.08	1.64	6.09	1.67	6.04	1.59	6.82	2.37	6.07	1.67	7.23	0.78	8.44	−0.03
1,3-C2B	9.10 ± 0.02	−[c]	7.13	1.13	6.16	1.70	6.17	1.73	6.12	1.64	6.90	2.43	6.15	1.73	7.32	0.83	8.51	0.03
1,4-C2B	8.92 ± 0.03	−[c]	6.95	1.16	5.96	1.74	5.97	1.77	5.92	1.68	6.70	2.47	5.97	1.76	7.12	0.84	8.31	0.05
1,2,3-C3B	9.18 ± 0.03	−[c]	7.29	1.30	6.29	1.82	6.29	1.85	6.25	1.77	7.02	2.55	6.29	1.87	7.46	1.03	8.68	0.21
1,2,4-C3B	9.04 ± 0.03	−[c]	7.11	1.40	6.09	1.94	6.10	1.97	6.05	1.89	6.83	2.68	6.12	1.98	7.27	1.10	8.46	0.31
1,3,5-C3B	9.30 ± 0.06	0.34	7.43	1.39	6.42	1.92	6.43	1.96	6.39	1.87	7.17	2.66	6.42	1.97	7.62	1.12	8.84	0.32
1,2,3,4-C4B	8.9	−[c]	7.23	1.52	6.18	2.02	6.20	2.06	6.15	1.97	6.93	2.77	6.23	2.11	7.37	1.24	8.58	0.45
1,2,3,5-C4B	9.02	−[c]	7.29	1.60	6.24	2.10	6.26	2.14	6.21	2.05	6.98	2.84	6.29	2.16	7.44	1.33	8.65	0.54
1,2,4,5-C4B	9	0.45	7.19	1.64	6.14	2.15	6.16	2.18	6.11	2.10	6.88	2.89	6.20	2.20	7.33	1.35	8.53	0.58
1,2,3,4,5-C5B	8.8	0.729	7.32	1.75	6.26	2.21	6.27	2.25	6.22	2.17	7.00	2.96	6.32	2.45	7.46	1.48	8.67	0.70
1,2,3,4,5,6-C6B	9.0 ± 0.1	0.92 ± 0.10	7.45	2.00	6.36	2.46	6.38	2.53	6.33	2.41	7.11	3.16	6.44	2.80	7.58	1.61	8.81	0.82
benzene	9.24	−1.12 ± 0.03[b]	7.01	0.42	6.19	1.05	6.18	1.07	6.15	0.99	6.94	1.77	6.04	1.04	7.26	0.11	8.43	−0.72

[a] The experimental *I* and *A* were obtained from NIST standard reference database webbook http://webbook.nist.gov/chemistry. [b] Ref 47. [c] Not listed in NIST database.

Table 6. Ionization Potential (*I*) and Electron Affinity (*A*) in eV Calculated using ΔSCF Method

| | B3LYP/6-31G* | | M05-2X/6-31++G** | | expt[a] | |
system	*I*	*A*	*I*	*A*	*I*	*A*
CB	8.68	−1.54	9.01	−1.09	9.07 ± 0.02	−[c]
1,2-C2B	8.69	−0.14	9.07	0.01	9.06 ± 0.02	0.093
1,3-C2B	8.75	−0.16	9.12	−0.60	9.10 ± 0.02	−[c]
1,4-C2B	8.56	−0.92	8.92	−0.58	8.92 ± 0.03	−[c]
1,2,3-C3B	8.86	0.17	9.24	0.29	9.18 ± 0.03	−[c]
1,2,4-C3B	8.65	0.16	9.04	0.27	9.04 ± 0.03	−[c]
1,3,5-C3B	9.00	0.15	9.39	−0.28	9.30 ± 0.06	0.34
1,2,3,4-C4B	8.73	0.44	9.12	0.52	8.9	−[c]
1,2,3,5-C4B	8.78	0.44	9.17	0.21	9.02	−[c]
1,2,4,5-C4B	8.66	0.07	9.08	0.05	9	0.45
1,2,3,4,5-C5B	8.76	0.70	9.19	0.75	8.8	0.729
1,2,3,4,5,6-C6B	8.86	0.77	9.29	0.87	9.0 ± 0.1	0.92 ± 0.10
benzene	8.86	−2.10	9.15	−1.48	9.24	−1.12 ± 0.03[b]

[a] The experimental *I* and *A* were obtained from NIST standard reference database webbook http://webbook.nist.gov/chemistry. [b] Ref 49. [c] Not listed in NIST database.

methods underestimate the ionization potentials. Further, the magnitudes and signs of the electron affinities have not been predicted properly by these methods. The values predicted from the M05-2X method are much closer to the experimental values. The discrepancy in the prediction of these values may be due to the parametrization of different density functionals and the limitations of Koopmans' theorem. In order to assess the importance of electron relaxation upon ionization and electron attachment, the ΔSCF approach are used employing M05-2X/6-31++G** method. The results are depicted in Table 6. It can be observed that ΔSCF values are in good agreement with the experimental values.

Global Descriptors of Chlorobenzene. The calculated chemical hardness, chemical potential, and electrophilicity index from various methods using 6-31++G** basis functions employing Koopmans' approximation and different

A QSAR Perspective

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2749**

Table 7. Global Reactivity Properties, Hardness (η), Chemical Potential (μ), and Electrophilicity (ω) in eV, Calculated from B3LYP, MPW1B95, MPWB1K, and BB1K Methods with 6-31++G** Basis Set using Koopmans' Theorem

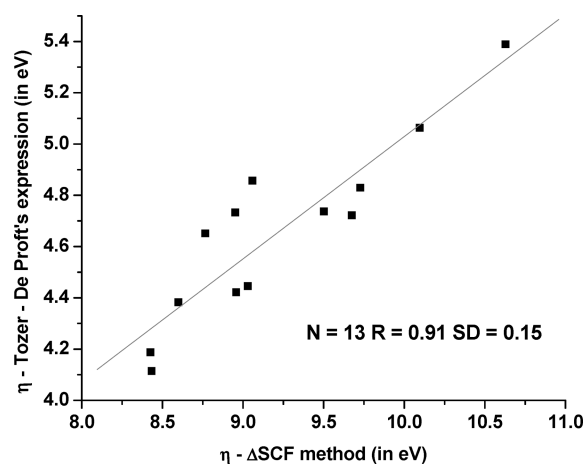| system	B3LYP			MPW1B95			MPWB1K			BB1K		
	η	μ	ω	η	μ	ω	η	μ	ω	η	μ	ω
CB	6.15	−3.88	1.22	4.59	−3.73	1.51	4.62	−3.71	1.49	4.63	−3.66	1.45
1,2-C2B	5.99	−4.07	1.39	4.42	−3.88	1.71	4.44	−3.86	1.68	4.46	−3.82	1.63
1,3-C2B	6.00	−4.13	1.42	4.44	−3.95	1.76	4.46	−3.93	1.73	4.48	−3.88	1.68
1,4-C2B	5.80	−4.05	1.42	4.20	−3.87	1.78	4.22	−3.85	1.76	4.24	−3.80	1.71
1,2,3-C3B	5.99	−4.29	1.54	4.44	−4.07	1.87	4.46	−4.05	1.84	4.48	−4.01	1.80
1,2,4-C3B	5.71	−4.26	1.59	4.13	−4.04	1.97	4.15	−4.01	1.94	4.17	−3.97	1.89
1,3,5-C3B	6.04	−4.41	1.61	4.47	−4.19	1.96	4.50	−4.17	1.93	4.52	−4.13	1.89
1,2,3,4-C4B	5.70	−4.37	1.68	4.14	−4.13	2.06	4.17	−4.10	2.02	4.17	−4.06	1.98
1,2,3,5-C4B	5.69	−4.45	1.74	4.12	−4.20	2.14	4.14	−4.17	2.10	4.16	−4.13	2.05
1,2,4,5-C4B	5.55	−4.42	1.76	3.97	−4.17	2.19	3.99	−4.15	2.16	4.01	−4.10	2.10
1,2,3,4,5-C5B	5.57	−4.54	1.85	4.02	−4.26	2.26	4.04	−4.23	2.22	4.05	−4.20	2.17
1,2,3,4,5,6-C6B	5.45	−4.72	2.05	3.85	−4.46	2.58	3.90	−4.41	2.50	3.92	−4.37	2.44
benzene	6.59	−3.72	1.05	5.11	−3.63	1.29	5.15	−3.62	1.27	5.16	−3.57	1.23

Table 8. Global Reactivity Properties, Hardness (η), Chemical Potential (μ), and Electrophilicity (ω) in eV, Calculated from SVWN5, BLYP, M05, and M05-2X Methods with 6-31++G** Basis Set using Koopmans' Theorem and Tozer-De Proft's Expression for Hardness

system	SVWN5			BLYP			M05			M05-2X			
	η	μ	ω	η	μ	ω	η	μ	ω	η	η^a	μ	ω
CB	4.63	−4.45	2.14	4.56	−3.69	1.50	6.66	−3.82	1.09	8.66	5.06	−4.01	0.93
1,2-C2B	4.45	−4.60	2.37	4.41	−3.87	1.70	6.46	−4.00	1.24	8.47	4.86	−4.20	1.04
1,3-C2B	4.47	−4.67	2.44	4.42	−3.94	1.75	6.49	−4.08	1.28	8.49	4.83	−4.27	1.07
1,4-C2B	4.23	−4.58	2.49	4.21	−3.87	1.77	6.28	−3.98	1.26	8.26	4.74	−4.18	1.06
1,2,3-C3B	4.47	−4.79	2.56	4.42	−4.08	1.88	6.44	−4.24	1.40	8.46	4.73	−4.45	1.17
1,2,4-C3B	4.15	−4.75	2.72	4.15	−4.05	1.98	6.17	−4.18	1.42	8.15	4.65	−4.39	1.18
1,3,5-C3B	4.51	−4.92	2.68	4.46	−4.20	1.98	6.50	−4.37	1.47	8.52	4.72	−4.58	1.23
1,2,3,4-C4B	4.16	−4.85	2.82	4.11	−4.17	2.11	6.13	−4.31	1.51	8.13	4.38	−4.52	1.25
1,2,3,5-C4B	4.14	−4.91	2.92	4.13	−4.23	2.16	6.11	−4.38	1.57	8.11	4.42	−4.60	1.30
1,2,4,5-C4B	3.99	−4.88	2.99	4.00	−4.20	2.21	5.98	−4.34	1.58	7.95	4.45	−4.55	1.30
1,2,3,4,5-C5B	4.04	−4.98	3.07	3.87	−4.38	2.48	5.98	−4.47	1.67	7.97	4.11	−4.69	1.38
1,2,3,4,5,6-C6B	3.95	−5.14	3.34	3.64	−4.62	2.93	5.97	−4.59	1.77	7.99	4.19	−4.81	1.45
benzene	5.17	−4.35	1.83	5.00	−3.54	1.25	7.15	−3.68	0.95	9.15	5.39	−3.85	0.81

[a] Tozer-De Proft's hardness expression using eq 10.
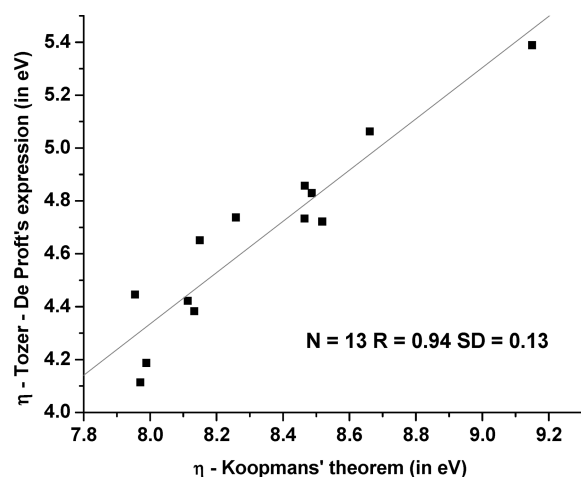
Figure 2. Plot between the hardness obtained from the Koopmans' theorem and the Tozer-De Proft's expression using M05-2X/6-31++G** method.

Figure 3. Plot between the hardness obtained from the ΔSCF method and the Tozer-De Proft's expression using M05-2X/6-31++G** method.

equations (eqs 3, 4, 7 and 10) are shown in Tables 7 and 8. The results obtained using the remaining basis functions are shown in Supporting Information (Tables S9−S19). The hardness of most of the CBs decreases with an increase in the number of chlorine atoms. The overall observation of the hardness profile indicates that for any of the chosen

method with different basis sets, the maximum and minimum hardness values lie within ~0.30 eV in most cases. The hardness predicted by MPW1B95, MPWB1K, BB1K, BLYP, and SVWN5 are in close agreement with each other. There is no such agreement in the values obtained from the other methods. The hardness and the relative energies of different CBs have been compared to understand the validity of
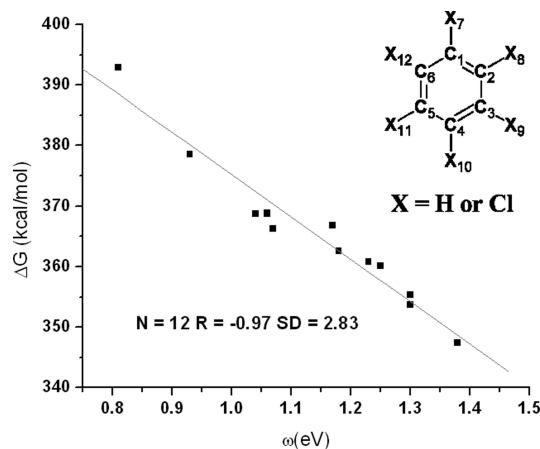
Table 9. Electrophilicity (ω), Calculated from M05-2X/ 6-31++G** Method, and Experimental Free Energy (ΔG_g^0 in kcal/mol) of Deprotonation of Benzene and CBs

system	ω(eV)	ΔG_g^0
CB	0.93	378.6
1,2-C2B	1.04	368.7
1,3-C2B	1.07	366.3
1,4-C2B	1.06	368.8
1,2,3-C3B	1.17	366.8
1,2,4-C3B	1.18	362.6
1,3,5-C3B	1.23	360.8
1,2,3,4-C4B	1.25	360.1
1,2,3,5-C4B	1.30	355.4
1,2,4,5-C4B	1.30	353.7
1,2,3,4,5-C5B	1.38	347.4
benzene	0.81	392.9

Figure 4. Plot between the electrophilicity (ω) calculated from M05-2X/6-31++G** method and the free energy (ΔG_g^0) of deprotonation.

maximum hardness principle (MHP). The stability (relative energies in Table 2) of dichloro systems vary as 1,4-C2B > 1,3-C2B > 1,2-C2B. The hardness calculated from all the methods applying Koopmans' approximation of these isomers vary as 1,3-C2B > 1,2-C2B > 1,4-C2B (Tables 7 and 8). In contrast, the hardness indices predicted from B3LYP/6-31G* method employing ΔSCF approach follow the order of 1,4-C2B > 1,3-C2B > 1,2-C2B (Supporting Information, Table S20), which is in parallel with the trend in the stabilities of these dichlorobenzenes. It can be noted that M05-2X/6-31G* method using ΔSCF does not yield the same trend. The 1,2,4-C3B isomer shows lowest hardness among the 1,2,3-C3B and 1,3,5-C3B isomers. The hardness of 1,2,4,5-C4B isomer of tetrachlorinated benzene marginally decreases when compared with the other tetrachlorinated benzenes. For each group of chlorobenzene isomers, in some cases, as the energy and electrophilicity decreases, the

hardness increases as expected from the principles of maximum hardness and minimum electrophilicity. These principles are not very effective in analyzing the relative stabilities of isomers.[3]

The hardness values calculated for benzene and CBs using Tozer and De Proft's expressions (eq 8) employing M05-2X/6-31++G** method are also given in Table 8. The calculated hardness values using Koopmans' approximation, ΔSCF method, and Tozer-De Proft's expression are compared in Figures 2 and 3. It can be seen that the relative trends in these values are similar.

The chemical potential and the electrophilicity index increase with an increase in the number of chlorine atom substitutions, which is in line with the results observed in

Table 10. Regression Analysis of Global Reactivity Descriptors, Hardness (η), Chemical Potential (μ), and Electrophilicity (ω), Calculated from Various Methods/Basis Set Using Koopmans' Theorem

method	basis set	η	μ	ω	method	basis set	η	μ	ω
		R					*R*		
B3LYP	6-31G*	−0.933	−0.887	0.949	SVWN5	6-31G*	−0.910	−0.862	0.966
	6-31G**	−0.929	−0.886	0.945		6-31G**	−0.903	−0.861	0.966
	6-31+G*	−0.909	−0.880	0.955		6-31+G*	−0.892	−0.848	0.971
	6-31+G**	−0.900	−0.880	0.960		6-31+G**	−0.887	−0.847	0.973
	6-31++G**	−0.900	−0.880	0.960		6-31++G**	−0.887	−0.846	0.973
	6-311++G**	−0.895	−0.877	0.959		6-311++G**	−0.877	−0.841	0.973
	AUG-CC-PVDZ	−0.892	−0.876	0.963					
	AUG-CC-PVTZ	−0.903	−0.868	0.957					
MPWB1K	6-31G*	−0.910	−0.868	0.961	BLYP	6-31G*	−0.921	−0.889	0.957
	6-31G**	−0.905	−0.867	0.961		6-31G**	−0.915	−0.889	0.961
	6-31+G*	−0.889	−0.855	0.969		6-31+G*	−0.897	−0.882	0.966
	6-31+G**	−0.883	−0.854	0.971		6-31+G**	−0.892	−0.882	0.969
	6-31++G**	−0.883	−0.854	0.971		6-31++G**	−0.893	−0.883	0.969
						6-311++G**	−0.888	−0.875	0.969
MPW1B95	6-31G*	−0.915	−0.871	0.960	M05	6-31G*	−0.945	−0.867	0.938
	6-31G**	−0.908	−0.870	0.960		6-31G**	−0.943	−0.864	0.934
	6-31+G*	−0.895	−0.859	0.966		6-31+G*	−0.930	−0.854	0.942
	6-31+G**	−0.888	−0.857	0.971		6-31+G**	−0.924	−0.854	0.946
	6-31++G**	−0.888	−0.857	0.971		6-31++G**	−0.927	−0.854	0.946
						6-311++G**	−0.916	−0.850	0.946
BB1K	6-31G*	−0.913	−0.870	0.961	M05-2X	6-31G*	−0.937	−0.869	0.931
	6-31G**	−0.905	−0.868	0.962		6-31G**	−0.936	−0.870	0.931
	6-31+G*	−0.895	−0.856	0.968		6-31+G*	−0.916	−0.861	0.938
	6-31+G**	−0.885	−0.855	0.968		6-31+G**	−0.912	−0.864	0.938
	6-31++G**	−0.885	−0.855	0.968		6-31++G**	−0.912	−0.864	0.938
						6-311++G**	−0.908	−0.861	0.944

Table 11. Regression Equations Obtained for ω from Various Methods/Basis Sets using Koopmans' Theorem

method	basis set	pC	R	SD	method	basis set	pC	R	SD
B3LYP	6-31G*	$(0.582)+3.248\omega$	0.949	0.219	SVWN5	6-31G*	$(-0.209)+2.204\omega$	0.966	0.178
	6-31G**	$(0.539)+3.259\omega$	0.945	0.227		6-31G**	$(-0.318)+2.237\omega$	0.966	0.180
	6-31+G*	$(-0.285)+3.337\omega$	0.955	0.205		6-31+G*	$(-0.976)+2.223\omega$	0.971	0.165
	6-31+G**	$(-0.409)+3.391\omega$	0.960	0.194		6-31+G**	$(-1.113)+2.258\omega$	0.973	0.160
	6-31++G**	$(-0.409)+3.391\omega$	0.960	0.194		6-31++G**	$(-1.113)+2.258\omega$	0.973	0.160
	6-311++G**	$(-0.505)+3.368\omega$	0.959	0.196		6-311++G**	$(-1.217)+2.228\omega$	0.973	0.161
	AUG-CC-PVDZ	$(-0.607)+3.401\omega$	0.963	0.186					
	AUG-CC-PVTZ	$(-0.571)+3.441\omega$	0.957	0.200					
MPWB1K	6-31G*	$(0.331)+2.868\omega$	0.961	0.191	BLYP	6-31G*	$(0.785)+2.553\omega$	-0.889	0.201
	6-31G**	$(0.302)+2.868\omega$	0.961	0.191		6-31G**	$(0.658)+2.617\omega$	-0.889	0.191
	6-31+G*	$(-0.446)+2.836\omega$	0.969	0.171		6-31+G*	$(-0.074)+2.577\omega$	-0.882	0.179
	6-31+G**	$(-0.589)+2.886\omega$	0.971	0.165		6-31+G**	$(-0.192)+2.619\omega$	-0.882	0.171
	6-31++G**	$(-0.589)+2.886\omega$	0.971	0.165		6-31++G**	$(-0.192)+2.619\omega$	-0.883	0.171
						6-311++G**	$(-0.352)+2.628\omega$	-0.875	0.170
MPW1B95	6-31G*	$(0.426)+2.762\omega$	0.960	0.195	M05	6-31G*	$(0.645)+3.476\omega$	-0.867	0.240
	6-31G**	$(0.315)+2.808\omega$	0.960	0.194		6-31G**	$(0.624)+3.481\omega$	-0.864	0.247
	6-31+G*	$(-0.427)+2.779\omega$	0.966	0.178		6-31+G*	$(-0.081)+3.558\omega$	-0.854	0.232
	6-31+G**	$(-0.547)+2.818\omega$	0.971	0.167		6-31+G**	$(-0.177)+3.607\omega$	-0.854	0.224
	6-31++G**	$(-0.547)+2.818\omega$	0.971	0.167		6-31++G**	$(-0.177)+3.607\omega$	-0.854	0.224
						6-311++G**	$(-0.386)+3.725\omega$	-0.850	0.225
BB1K	6-31G*	$(0.377)+2.854\omega$	-0.870	0.190	M05-2X	6-31G*	$(0.454)+4.328\omega$	-0.869	0.253
	6-31G**	$(0.263)+2.903\omega$	-0.868	0.190		6-31G**	$(0.454)+4.328\omega$	-0.870	0.253
	6-31+G*	$(-0.407)+2.886\omega$	-0.856	0.175		6-31+G*	$(-0.539)+4.639\omega$	-0.861	0.239
	6-31+G**	$(-0.497)+2.911\omega$	-0.855	0.173		6-31+G**	$(-0.539)+4.639\omega$	-0.864	0.239
	6-31++G**	$(-0.497)+2.911\omega$	-0.855	0.173		6-31++G**	$(-0.539)+4.639\omega$	-0.864	0.239
						6-311++G**	$(-0.689)+4.618\omega$	-0.861	0.228

the case of hardness. The chemical potential predicted by MPW1B95, MPWB1K, BB1K, BLYP, and M05 are similar. There is only a marginal variation in the electrophilicity index predicted by the MPW1B95, MPWB1K, and BB1K methods.

Relation Between Recativities and Electrophilicity. With a view to assess how electrophilicity index predicts the reactivity of a series of molecules, the relationship between the deprotonation free energies[39] of benzene and chlorobenzenes (Table 9) and the electrophilicity has been explored, and it is depicted in Figure 4. It can be seen that deprotonation free energies vary linearly with electrophilicity, and thus, electrophilicity provides valuable information about the relative reactivities of a series of homologous compounds.

Structure-Toxicity Analysis of Chlorobenzenes. With a view to benchmark the level of calculation of different global reactivity descriptors for QSTR development, the toxicity (pC) values of the selected set of nine chlorobenzenes against *Poecilia reticulata* have been used. In the development of QSTR, the experimental toxicity data (pC) have been considered as dependent variables and global reactivity descriptors obtained from all the methods (Tables 7−8, and Supporting Information, Tables S9−S19) as independent variables. The QSTRs derived from the multiple regression analysis employing different global reactivity descriptors obtained from various levels of calculations are presented in Tables 10−11.

The ΔSCF method has also been used along with the Koopmans' approximation to understand the predictive power of global reactivity descriptors obtained from different strategies. The eqs 5 and 6 have been used to predict the ionization potential and the electron affinity using B3LYP and M05-2X methods with 6-31G* and 6-31++G** basis sets, respectively. The predicted regression values using ΔSCF method is shown in the Table 12. The experimental

Table 12. Regression Analysis of Global Reactivity Descriptors, Hardness (η), Chemical Potential (μ), and Electrophilicity (ω), Calculated from Various Methods/Basis sets using ΔSCF Method

method/basis set	η		μ		ω	
	R	SD	R	SD	R	SD
B3LYP/6-31G*	-0.783	0.431	-0.693	0.499	0.734	0.470
M05-2X/6-31++G**	-0.738	0.467	-0.768	0.443	0.768	0.443

and calculated toxicity data (pC) for the selected set of nine CBs are listed in Tables 13 and 14.

As shown in previous studies, electrophilicity index exhibits very good correlation with the toxicity values when compared to hardness and electronegativity.[40] Even though, the global reactivity descriptors calculated using different methods show significant variations as discussed above, the mean deviation of the properties between all the systems with respect to the chosen method/basis set is similar. That is the predicted global reactivity descriptors are varying with respect to the method and the basis set, but the magnitude of the difference in between the systems are similar in all the methods. This observation influences the final regression analysis and predicts the correlation coefficient without major deviation between each of the methods. Although the ionization potential and the electron affinity calculated using ΔSCF methods produce good correlation with the experimentally observed data, they do not improve the correlation coefficient.

In general, the regression values obtained using various methods are comparable. The overall observation illustrates that the increasing the number of basis functions in any method refines the calculated values of the descriptors only. There is no marked variation in the correlation coefficient. It is an interesting observation that quantum chemical descriptors such as the global reactivity descriptors can be routinely used in the

Table 13. Experimental and Calculated Fish Toxicity (*pC*) Values against *Poecilia Reticulate* of Selected Chlorobenzenes from B3LYP, MPW1B95, MPWB1K, and BB1K Methods with 6-31++G** Basis Set using Koopmans' Theorem

system	observed	B3LYP			MPW1B95			MPWB1K			BB1K		
		ω (eV)	calc	res	ω (eV)	calc	res	ω (eV)	calc	res	ω (eV)	calc	res
CB	3.77	1.22	3.73	0.04	1.51	3.72	0.05	1.49	3.72	0.05	1.45	3.72	0.05
1,2-C2B	4.4	1.39	4.29	0.11	1.71	4.26	0.14	1.68	4.26	0.14	1.63	4.26	0.14
1,3-C2B	4.28	1.42	4.41	−0.13	1.76	4.40	−0.12	1.73	4.40	−0.12	1.68	4.40	−0.12
1,4-C2B	4.56	1.42	4.40	0.16	1.78	4.47	0.09	1.76	4.48	0.08	1.71	4.47	0.09
1,2,4-C3B	4.83	1.59	4.97	−0.14	1.97	5.01	−0.18	1.94	5.00	−0.17	1.89	5.01	−0.18
1,3,5-C3B	4.74	1.61	5.06	−0.32	1.96	4.99	−0.25	1.93	4.99	−0.25	1.89	5.00	−0.26
1,2,3,4-C4B	5.35	1.68	5.28	0.07	2.06	5.25	0.10	2.02	5.24	0.11	1.98	5.26	0.09
1,2,3,5-C4B	5.43	1.74	5.49	−0.06	2.14	5.48	−0.05	2.10	5.48	−0.05	2.05	5.48	−0.05
1,2,4,5-C4B	5.85	1.76	5.55	0.30	2.19	5.62	0.23	2.16	5.63	0.22	2.10	5.61	0.24

Table 14. Experimental and Calculated Fish Toxicity (*pC*) Values against *Poecilia Reticulate* of Selected Chlorobenzenes from SVWN5, BLYP, M05, and M05-2X Methods With 6-31++G** Basis Set using Koopmans' Theorem

system	observed	SVWN5			BLYP			M05			M05-2x		
		ω (eV)	calc	res	ω (eV)	calc	res	ω (eV)	calc	res	ω (eV)	calc	res
CB	3.77	2.14	3.71	0.06	1.50	3.73	0.04	1.09	3.77	0.00	0.93	3.76	0.01
1,2-C2B	4.4	2.37	4.25	0.15	1.70	4.26	0.14	1.24	4.30	0.10	1.04	4.30	0.10
1,3-C2B	4.28	2.44	4.39	−0.11	1.75	4.40	−0.12	1.28	4.44	−0.16	1.07	4.45	−0.17
1,4-C2B	4.56	2.49	4.50	0.06	1.77	4.45	0.11	1.26	4.38	0.18	1.06	4.38	0.18
1,2,4-C3B	4.83	2.72	5.03	−0.20	1.98	4.99	−0.16	1.42	4.94	−0.11	1.18	4.94	−0.11
1,3,5-C3B	4.74	2.68	4.95	−0.21	1.98	4.98	−0.24	1.47	5.13	−0.39	1.23	5.17	−0.43
1,2,3,4-C4B	5.35	2.82	5.26	0.09	2.11	5.35	0.00	1.51	5.29	0.06	1.25	5.28	0.07
1,2,3,5-C4B	5.43	2.92	5.47	−0.04	2.16	5.47	−0.04	1.57	5.49	−0.06	1.30	5.50	−0.07
1,2,4,5-C4B	5.85	2.99	5.63	0.22	2.21	5.59	0.26	1.58	5.50	0.35	1.30	5.51	0.34

development of QSAR/QSPR/QSTR approaches in the industries without any high computational time and cost.

Conclusions

In this study, an attempt has been made to evaluate how various DFT methods predict the stability, the electronic properties, and the reactivities of CBs. In this investigation, some popular functionals belonging to the pure, GGA, hybrid, and meta and hybrid meta GGA have been selected. It is evident from the results that increase in the rigor in the level of calculation leads to good agreement between the calculated and experimental values of ionization potentials and electron affinities. Typically, M05-2X method outperforms the other functionals in the prediction of ionization potentials and electron affinities of CBs using Koopmans' approximation and ΔSCF method. The ionization potential and the electron affinities calculated from ΔSCF method are in good agreement with the experimental values when compared to the Koopmans' approximation. The trends in the hardness values of CBs calculated from various schemes are comparable.

The usefulness of global reactivity descriptors in the development of structure−activity/toxicity relationships has been illustrated. There is no drastic change in the correlation coefficients obtained from the linear/multiple regression analysis of toxicity vs calculated global reactivity descriptors from different levels of theory. The results clearly show that the reliable estimates of the trends in the global reactivity descriptors are sufficient enough to develop QSAR/QSPR/QSTR. This investigation, once again, reinforces the usefulness of global reactivity descriptors in QSAR/QSPR/QSTR parlance.

Supporting Information Available: Provides the calculated bond lengths and bond angles of chlorobenzene and 1,3,5-C3B from various methods using 6-31++G** basis set, the calculated ionization potential and electron affinity using various methods/basis sets, and the various global reactivity descriptors like hardness, chemical potential, electrophilicity index, etc., calculated using various methods/basis sets. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chermette, H. *J. Comput. Chem.* **1999**, *20*, 129.

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

(3) Pearson, R. G. *Chemical Hardness: Applications from Molecules to Solids*; VCH-Wiley: Weinheim, Germany, 1997.

(4) Tozer, D. J.; De Proft, F. *J. Phys. Chem. A* **2005**, *109*, 8923.

(5) De Proft, F.; Sablon, N.; Tozer, D. J.; Geerlings, P. *Faraday Discuss.* **2007**, *135*, 151.

(6) (a) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618. (b) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221. (c) Paizs, B.; Suhai, S. *J. Comput. Chem.* **1998**, *19*, 575. (d) Alkorta, I.; Rozas, I.; Elguero, J. *J. Phys. Chem. A* **1998**, *102*, 9278. (e) Rappe, A. K.; Bernstein, E. R. *J. Phys. Chem. A* **2000**, *104*, 6117. (f) Karpfen, A. *J. Phys. Chem. A* **2001**, *105*, 2064.

A QSAR Perspective

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2753**

(7) (a) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, 864. (b) Kohn, W.; Sham, L.*J. Phys. Rev.* **1965**, *140*, 1133. (c) Perdew, J. P.; Zunger, A. *Phys. Rev. B: Condens. Matter* **1981**, *23*, 5048. (d) Perdew, J. P. *Phys. Rev. B: Condens. Matter* **1986**, *33*, 8822. (e) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098. (f) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785. (g) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

(8) Yan Zhao, Y.; Gonza͞lez-Garć, N. R.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.

(9) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(10) Zhao, Y.; Pu, J.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2004**, *6*, 673.

(11) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.

(12) (a) Parthasarathi, R.; Padmanabhan, J.; Subramanian, V.; Maiti, B.; Chattaraj, P. K. *J. Phys. Chem. A* **2003**, *107*, 10346. (b) Melin, J.; Aparicio, F.; Subramanian, V.; Galva'n, M.; Chattaraj, P. K. *J. Phys. Chem. A* **2004**, *108*, 2487. (c) Parthasarathi, R.; Subramanian, V.; Roy, D. R.; Chattaraj, P. K. *Bioorg. Med. Chem.* **2004**, *12*, 5533. (d) Roy, D. R.; Parthasarathi, R.; Padmanabhan, J.; Sarkar, U.; Subramanian, V.; Chattaraj, P. K. *J. Phys. Chem. A* **2006**, *110*, 1084. (e) Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. *J. Phys. Chem. A* **2006**, *110*, 2739. (f) Gao, A. *J. Mol. Struct. (THEOCHEM)* **2007**, *816*, 67.

(13) (a) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524. (b) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **199677**, 3865. (c) Schultz, N.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127. Quintal, M. M.; Karton, A.; Iron, M. A.; Boese, A. D.; Martin, J. M. L. *J. Phys. Chem. A* **2006**, *110*, 709.

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

(15) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(16) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.

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

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

(19) Perdew, J. P.; Schmidt, K. *In Density Functional Theory and Its Applications to Materials*; Doren, V., Alsenoy, C. V., Geerlings., P., Eds.; American Institute of Physics: New York, 2001.

(20) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.

(21) Rose, K.; Hall, L. H. *SAR QSAR Environ. Res.* **2003**, *14*, 113.

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

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

(24) Pearson, R. G. *J. Am. Chem. Soc.* **1963**, *85*, 3533.

(25) Pearson, R. G. *Acc. Chem. Res.* **1993**, *26*, 250.

(26) Parr, R. G.; Chattaraj, P. K. *J. Am. Chem. Soc.* **1991**, *113*, 1854.

(27) Chattaraj, P. K.; Liu, G. H.; Parr, R. G. *Chem. Phys. Lett.* **1995**, *237*, 171.

(28) Chattaraj, P. K. *Proc. Indian Natl. Sci. Acad., Part A* **1996**, *62*, 513.

(29) Chattaraj, P. K.; Fuentealba, P.; Gomez, B.; Contreras, R. *J. Am. Chem. Soc.* **2000**, *122*, 348.

(30) Iczkowski, R. P.; Margrave, J. L. *J. Am. Chem. Soc.* **1961**, *83*, 3547.

(31) Putz, M. V.; Russo, N.; Sicilia, E. *J. Comput. Chem.* **2004**, *25*, 994.

(32) Mineva, T.; Heine, T. *J. Phys. Chem. A* **2004**, *108*, 11086.

(33) Zhan, C. G.; Nichols, J. A.; Dixon, D. A. *J. Phys. Chem. A* **2003**, *107*, 4184.

(34) Parr, R. G.; Szentpaly, L. v.; Liu, S. *J. Am. Chem. Soc.* **1999121**, 1922.

(35) (a) Chattaraj, P. K.; Sarkar, U.; Roy, D. R. *Chem. Rev.* **2006**, *106*, 2065. (b) Chattaraj, P. K.; Roy, D. R. *Chem. Rev.* **2007**, *107*, PR46.

(36) 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 E.01; Gaussian, Inc.: Wallingford, CT, 2004.

(37) Cradock, S.; Muir, J. M.; Rankin, D. W. H. *J. Mol. Struct.* **1990**, *220*, 205.

(38) Blom, R.; Cradock, S.; Davidson, S. L.; Rankin, D. W. H. *J. Mol. Struct.* **1991**, *245*, 369.

(39) Schlosser, M.; Marzi, E.; Cottet, F.; Buker, H. H.; Nibbering, N. M. M. *Chem.—Eur. J.* **2001**, *7*, 3511.

(40) Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. *J. Phys. Chem. A* **2005**, *109*, 11043.

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

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

(43) Hertwig, R. H.; Koch, W. *Chem. Phys. Lett.* **1997**, *268*, 345.

(44) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(45) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785.

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

(47) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

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

(49) Burrow, P. D.; Michejda, J. A.; Jordan, K. D. *J. Chem. Phys.* **1987**, *86*, 9.

JCTC Journal of Chemical Theory and Computation

Investigation of Exchange Energy Density Functional Accuracy for Interacting Molecules

Éamonn D. Murray,* Kyuho Lee, and David C. Langreth

Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, New Jersey 08854-8019

Abstract: We present a comparison of exchange-only interaction energies obtained using several standard exchange functionals in the generalized gradient approximation to Hartree−Fock results for interacting molecules. We observe that functionals with an enhancement factor using a 2/5 power dependence on the gradient of the density for large density gradients offer consistently better agreement with Hartree−Fock calculations than that of alternative functionals. We revisit the functional offering the closest agreement and recalculate it to include its exact large gradient dependence.

There has recently been a large interest and some progress in the treatment of van der Waals or dispersion interactions within or as an adjunct to density functional theory (DFT) calculations. These include empirical functionals,[1−4] a collection of methods of the DFT-D type,[5−9] a modified atomic potential method,[10−14] the Becke−Johnson method,[15−22] a nonempirical van der Waals density functional (vdW-DF)[23−28] and variants,[29] and a number of others.[30−43] The development and application of such methods are beginning to allow the application of density functional theory to new classes of matter and are of great importance. Some of these methods calculate the exchange energy within Hartree−Fock (HF), while others use some form of generalized gradient approximation (GGA). For the former, as the DFT methods for including dispersion become more efficient, one may look for a GGA method for replacing the less efficient HF part of the calculation. For the latter, one may also wish to find the GGA for exchange which best mimics HF. The evaluation of some of the common GGAs for exchange in comparison with HF is an aim of this work. As pointed out by Lacks and Gordon,[44] the functionals which are most accurate out on the long-range tail of the wave function and so important for van der Waals complexes may be quite different from those which are most successful for isolated molecules or dense condensed matter. A systematic study of the performance of exchange functionals for such systems of interest will be detailed here.

One important issue which is not addressed here is the compatibility of a particular exchange functional and the corresponding approximation for correlation. It is well-known that one of the reasons for the great success of DFT is the cancellation of errors between exchange and correlation. Each can have long-range parts which cancel in their sum, a fact that can be especially important for developing approximations for open shell systems. The sole criterion used in this work is how well each of the tested GGA exchange functionals reproduce the results of HF. The user will have to judge whether that is the correct criterion for the matching correlation approximation to be used.

We begin this paper with a brief description of the issue of spurious exchange binding and how it may be overcome by the generalized gradient approximation (GGA). This is followed by a summary of the exchange functionals we will be using in our calculations. These exchange functionals are: PBE,[45] revPBE,[46] PW86,[47] Becke86b,[48] and Becke88.[49] This is followed by a comparison of the calculated interaction energy as a function of molecular separation for pairs of molecules of various sizes. These calculations are performed using both HF and exchange-only density functional theory. The systems we examine are dimers consisting of pairs of the following atoms: H_2, N_2, CO_2, ammonia, methane, ethene, benzene and pyrazine. We conclude the paper with a re-examination of the PW86 functional, which, as we will show, offers the most consistent agreement with HF interaction energies. This work follows the pioneering work of Lacks and Gordon[44] and the recent work of Kannemann and

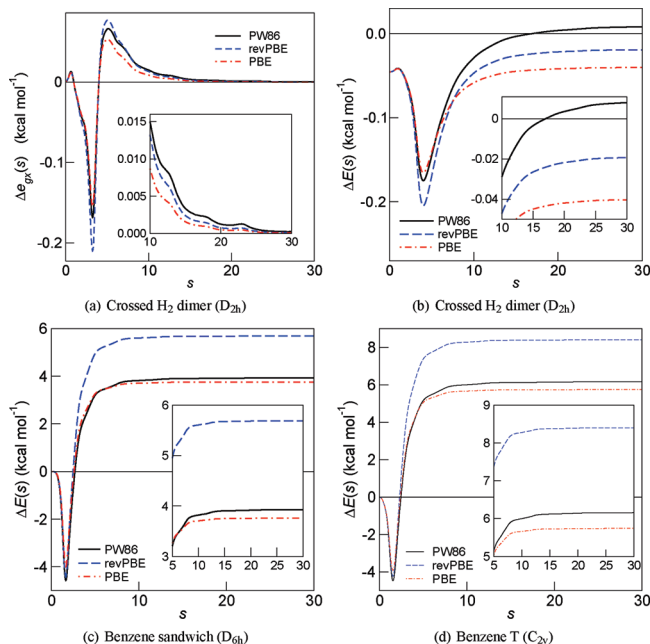* Corresponding author. E-mail: emurray@physics.rutgers.edu.

Exchange-Energy Density Functional Accuracy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2755**

Figure 1. Large s contribution to exchange repulsion of (a−b) crossed H_2 dimer 4 Å apart, (c) parallel benzene molecules 3.75 Å apart, and (d) benzene molecules in the T configuration 5.0 Å apart. (a) The differential exchange energy of crossed H_2 dimer showing the *explicitly* gradient-dependent contribution to the exchange interaction energy from each s value. (b) The interaction energy $\Delta E(s)$ of the crossed H_2 dimer as a function of s, showing the convergence of the exchange energy as a function of s. Results obtained from PW86, revPBE, and PBE are compared. Note that, of these, only PW86 gives correct repulsive interaction (positive ΔE). (c−d) The interaction energy $\Delta E(s)$ of parallel benzene molecules and benzene molecules in the T configuration as a function of s, respectively.
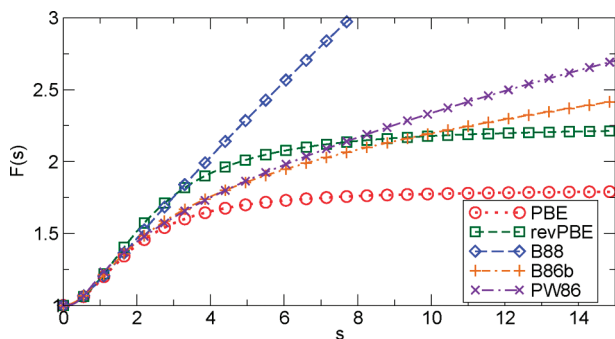
Figure 2. Enhancement factors $F(s)$ plotted for the exchange functionals discussed in this paper.

Becke,[50] who have come to a similar conclusion for the case of interacting rare gas atoms.

All of the exchange functionals discussed in this paper use the generalized gradient approximation (GGA) to exchange, so can be written in terms of an enhancement factor $F(s)$, where the exchange energy of a system with density $n(\mathbf{r})$ is given by

$$E_x^{\text{GGA}}[n(\mathbf{r}), \nabla n(\mathbf{r})] = A_x \int d^3 r\, n^{4/3} F_x^{\text{GGA}}(s) \quad (1)$$

with $s = (\nabla n)/(2k_F n)$, where $k_F \equiv (3\pi^2 n)^{1/3}$ is the local Fermi wavevector, and $A_x = -(3/4)(3/\pi)^{1/3}$.

The issue of this spurious exchange binding has been studied previously by Harris[51] in the case of two interacting He atoms. In examining the system in terms of the first-order change in the potential due to the overlap in the densities:

$$M = \int d^3 r\, \psi_{\text{He}}(\mathbf{r} - \mathbf{R}_1)\psi_{\text{He}}(\mathbf{r} - \mathbf{R}_2) \quad (2)$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ are the nuclear locations and

$$\psi_{\text{He}}(\mathbf{r}) = \frac{1}{\sqrt{2}} n^{1/2}(\mathbf{r}) \quad (3)$$

He noted that the leading contribution to the change in the energy comes from competition between the kinetic and the exchange−correlation terms. In HF, both the kinetic term, which is repulsive, and the exchange term, which is attractive, are proportional to M^2. This yields an interaction energy which is monotonic in $|\mathbf{R}_1 - \mathbf{R}_2|$ and always repulsive, and which may be written schematically as

$$\delta E \simeq (C_T^{\text{HF}} - C_x^{\text{HF}})M^2 \quad (4)$$

However, in the case of LDA, the exchange term is proportional to $M^{4/3}$, leading to a situation where LDA can yield an attractive interaction for small overlaps (small values of M), so that eq 4 is replaced by

$$\delta E \simeq C_T M^2 - C_x M^{4/3} \quad (5)$$

The second term of eq 5 dominates at large separations and gives spurious binding from exchange alone. This shortcoming of the LDA is generally counteracted to an extent by GGAs due to the enhancement factor, increasing the exchange contribution in regions of small density overlap. This is because an increasing density overlap decreases the value of s in the overlap region, thus reducing the enhancement of a negative contribution to the energy and thereby giving repulsion, which can compensate for the LDA's spurious attraction. However in a highly inhomogeneous region, a GGA functional whose enhancement factor becomes flat for large s will give an exchange energy with the same dependence on the overlap as in LDA, so that the LDA's spurious exchange binding will be exacerbated instead of ameliorated in regions of inhomogeneous low density.

In order to illustrate the importance of the asymptotic behavior of the enhancement factor $F_x(s)$ for large s, we analyze the gradient exchange energy contribution from each s for the case of the crossed H_2 dimer with the PBE, revPBE, and PW86 exchange functionals. In Figure 1(a), we show the *explicitly* gradient-dependent component of the differential exchange interaction energy:

$$\Delta e_{\text{gx}}(s) = e_{\text{gx}}^{2H_2}(s) - 2e_{\text{gx}}^{H_2}(s) \quad (6)$$

where the gradient component of the differential exchange energy $e_{\text{gx}}(s)$ is defined by

$$e_{\text{gx}}(s) = A_x \int d^3 r\, n^{4/3}[F_x(s(\mathbf{r})) - 1]\delta(s - s(\mathbf{r})) \quad (7)$$
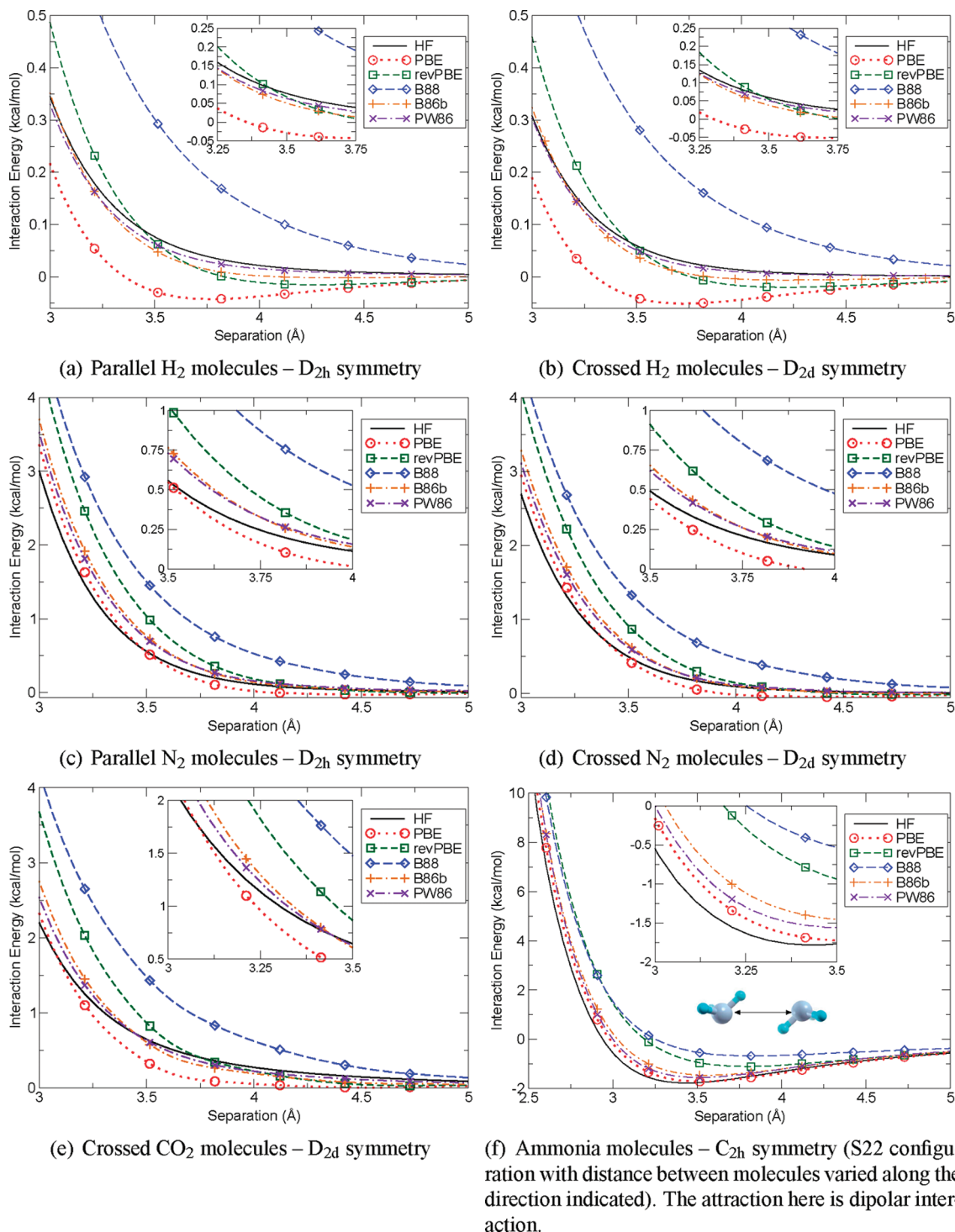
(a) Parallel H_2 molecules – D_{2h} symmetry

(b) Crossed H_2 molecules – D_{2d} symmetry

(c) Parallel N_2 molecules – D_{2h} symmetry

(d) Crossed N_2 molecules – D_{2d} symmetry

(e) Crossed CO_2 molecules – D_{2d} symmetry

(f) Ammonia molecules – C_{2h} symmetry (S22 configuration with distance between molecules varied along the direction indicated). The attraction here is dipolar interaction.
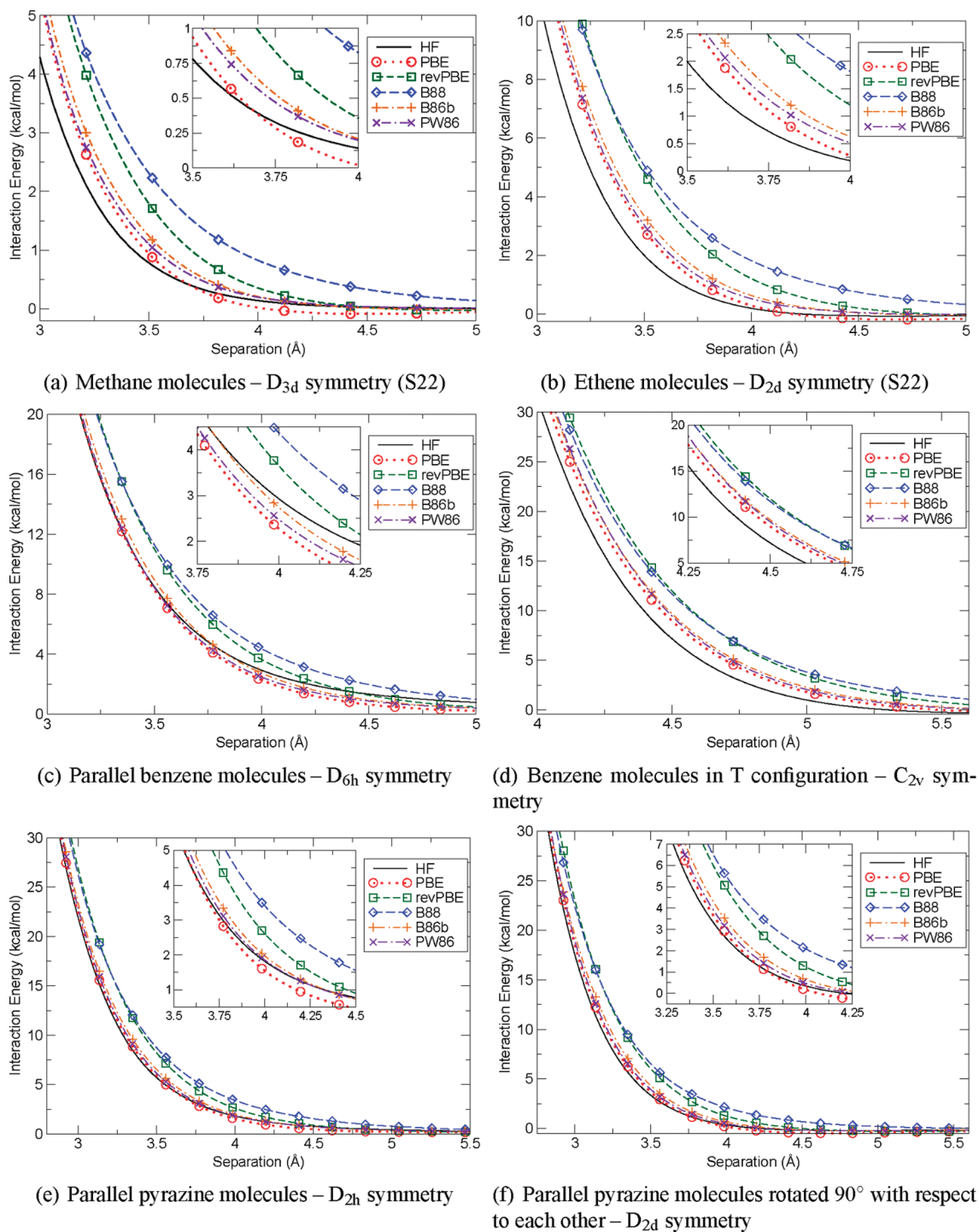
Figure 3. DFT exchange-only interaction energies calculated using several different exchange functionals compared to HF results for pairs of interacting molecules. Those structures marked as S22 use geometries taken from ref 58. The separations refer to the distance between the centers of mass of the molecules.

The term $\delta(s - s(\mathbf{r}))$ has been employed previously in ref 52. Note that in eq 7 the larger LDA component of the differential exchange energy:

$$e_{\text{LDA}x}(s) = A_x \int d^3r\, n^{4/3} \delta(s - s(\mathbf{r})) \qquad (8)$$

has been explicitly subtracted out, so that the difference between each GGA flavor is more visible. Figure 1(a) shows the contribution of gradient exchange energy $\Delta e_{\text{gx}}(s)$ to the interaction energy from each value of s.

By integrating $\Delta e_{\text{gx}}(s)$ over s and adding all the remaining energy terms ΔE_R, the interaction energy ΔE as a function of s is obtained, i.e.,

$$\Delta E(s) = \int_0^s ds' \Delta e_{\text{gx}}(s') + \Delta E_R \qquad (9)$$

Note that ΔE_R also contains the LDA exchange energy contribution arising from $\int_0^\infty ds\, e_{\text{LDA}x}(s)$. The convergence of interaction energy ΔE as a function of s is shown in

Exchange-Energy Density Functional Accuracy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2757**

(a) Methane molecules – D_{3d} symmetry (S22)

(b) Ethene molecules – D_{2d} symmetry (S22)

(c) Parallel benzene molecules – D_{6h} symmetry

(d) Benzene molecules in T configuration – C_{2v} symmetry

(e) Parallel pyrazine molecules – D_{2h} symmetry

(f) Parallel pyrazine molecules rotated 90° with respect to each other – D_{2d} symmetry

Figure 4. DFT exchange-only interaction energies calculated using several different exchange functionals compared to HF results for pairs of interacting molecules. Those structures marked as S22 use geometries taken from ref 58. The separations refer to the distance between the centers of mass of the molecules.

Figure 1(b). Note that $\Delta E(\infty)$ is equal to the total interaction energy.

From this it is clear that there is still a significant contribution for s as large as 25 in the H_2 dimer cases and 15 in the benzene cases. We also note that, as will be shown in our later calculations, the most consistent agreement with HF results is obtained from functionals with an enhancement factor proportional to $s^{2/5}$ at large s. Becke88, which is almost linear in s for large s, yields overly repulsive interaction energies in all cases. PBE and revPBE, which are constructed

to tend to a constant as s is increased yield spurious attractions in several cases.

The PBE and revPBE functionals use a form originally proposed by Becke[53] but with different parameters. The corresponding enhancement factor can be written as

$$F_x(s) = 1 + \frac{\mu s^2}{1 + \mu s^2/\kappa} \tag{10}$$

where both PBE and revPBE set $\mu = 0.2195$ to recover the expected low s limit. In PBE, κ is fixed to be 0.804 so the

local Lieb−Oxford bound is satisfied. This guarantees that the global Lieb−Oxford bound[54] will be satisfied for any possible density, even if the functional is not valid at that density. On the other hand, the other functionals discussed here may exceed the global bound for densities for which they are invalid. In the revPBE functional, this constraint is relaxed, and $\kappa = 1.245$ is obtained by fitting to atomic total energies. The PW86 enhancement factor is given by

$$F_x(s) = (1 + 15as^2 + bs^4 + cs^6)^{1/15} \qquad (11)$$

with $a = 0.0864$, $b = 14$, and $c = 0.2$. The value of a was fixed to recover the then expected low-s dependence while b and c are obtained from a fit to the numerical solution of an integral derived from the gradient expansion of the exchange hole with real space cutoffs. The full derivation is discussed in more detail toward the end of this paper. The Becke86b functional was derived by considering highly inhomogeneous systems where it was shown that the large-s dependence of the enhancement factor should be $s^{2/5}$. Its final form was chosen so as to smoothly interpolate between this large s form and the expected small s form. It can be cast into a PBE-like enhancement factor as

$$F_x(s) = 1 + \frac{\mu s^2}{(1 + \mu s^2/\kappa)^{4/5}} \qquad (12)$$

where $\mu = 0.2449$ and $\kappa = 0.5757$ are obtained by fitting to noble gas exchange energies. Finally, the Becke88 functional is constructed so as to reproduce the exact asymptotic behavior of a finite many electron system: $\lim_{r \to \infty} U_x^\sigma = -1/r$ with U_x^σ the Coulomb potential of exchange charge density $\rho_x^\sigma(r,r')$. One can also write this in a PBE-like form as

$$F_x(s) = 1 + \frac{\mu s^2}{1 + \frac{9}{4\pi}\mu s \sinh^{-1}(\lambda s)} \qquad (13)$$

where $\lambda = 2(6\pi^2)^{1/3}$, and $\mu = 0.2743$ is found by fitting to noble gas exchange energies. $F(s)$ for each of these functionals is plotted in Figure 2.

Using each of these functionals, we calculate the exchange-only interaction as a function of separation for pairs of molecules. These calculations are performed using abinit,[55] a plane wave density functional theory code, and a 25 Ha
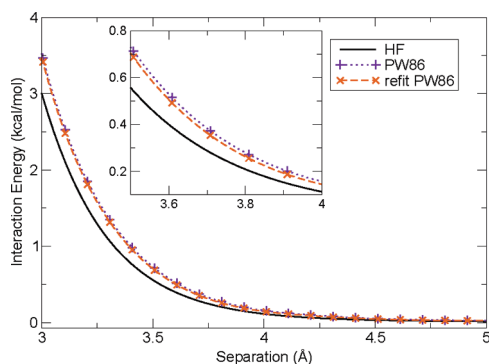

Figure 5. The exchange-only interaction energy for parallel N_2 molecules calculated using both the original and the refit PW86 functionals along with the HF results.

plane wave energy cutoff with the molecules placed in a 30 × 30 × 30 Bohr box (this is increased to 30 × 30 × 40 Bohr for the benzene and pyrazine systems). The corresponding HF calculations are performed using gamess.[56,57] In all of these calculations, the basis set superposition error is counterpoise corrected. All the systems are calculated using the aug-cc-pVTZ basis set with the exception of the benzene and the pyrazine systems, where the aug-cc-pVDZ basis is used. Testing indicates the error in the HF interaction due to use of the smaller basis set is ≲ 1% (see Supporting Information). In each case, we calculate the self-consistent interaction energy for the respective GGA functional and for HF. Because of the importance of the cancellation between kinetic and exchange energies discussed above and hence, preserving it in our comparisons, we believe this procedure is more relevant than a comparison of either exchange energies alone or post process calculations.

The results of these calculations are plotted in Figures 3 and 4. It is clear from the figures that both PW86 and Becke86b offer, for the most part, better agreement with HF than the other functionals, with PW86 performing slightly better. Both of these functionals are proportional to $s^{2/5}$ for large s, with $F_x^{PW86}(s) \sim 0.898s^{2/5}$ and $F_x^{B86b}(s) \sim 0.485s^{2/5}$. A similar conclusion has been reached in previous studies of rare gas diatomics by Lacks and Gordon[44] and more recently by Kannemann and Becke.[50] Here, we find that, while PW86 and Becke86b are consistently the best for systems dominated by large s, we found several cases where PBE was marginally better in the binding region. The error in the interaction energy ($Ex_{GGA} - E_{HF}$) for PBE, revPBE, PW86, and B86b compared to the accurate CCSD(T) calculations of the full binding energy at both the optimal separation and larger separations is shown in Table 1 for several systems. This gives a good indication of the effect on the binding energy of changing the GGA used in the calculation. It is also important to note that, as can be seen in the table, the PW86 functional is the only one of those tested that did not yield spurious exchange binding in any of the systems examined.

Of primary importance is the question of the true large s dependence of the integrals solved to obtain the PW86 functional. These are obtained from the expression for the gradient expansion of the exchange hole density:

$$n_x^{GEA}(\mathbf{r}, \mathbf{r} + \mathbf{R}) = -\frac{1}{2}n(\mathbf{r})y(\mathbf{r}, \mathbf{R}) \qquad (14)$$

where

$$y = J + 4L\hat{\mathbf{R}} \cdot \mathbf{s}/3 - 16M(\hat{\mathbf{R}} \cdot \mathbf{s})^2/27 - 16Ns^2/3 \qquad (15)$$

with

$$J = 72[4 + z^2 - (4 - z^2)\cos z - 4z \sin z]z^6 \qquad (16)$$

Exchange-Energy Density Functional Accuracy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2759**

$$L = 9(2 - 2\cos z - z\sin z)/z^3 \qquad (17)$$

$$M = 9(-z\cos z + \sin z)/(16z) \qquad (18)$$

$$N = 3[8 - (8 - 4z^2)\cos z - (8z - z^3)\sin z]/(16z^4) \qquad (19)$$

and $z = 2k_F R$. In ref 47, it was observed that while n_x^{GEA} violates the following two conditions for exchange hole:

$$n_x(\mathbf{r}, \mathbf{r} + \mathbf{R}) \leq 0 \qquad (20)$$

and

$$\int d^3 R n_x(\mathbf{r}, \mathbf{r} + \mathbf{R}) = -1 \qquad (21)$$

these could be satisfied by applying cutoffs so that:

$$n_x^{GGA}(\mathbf{r}, \mathbf{r} + \mathbf{R}) = -\frac{1}{2}n(\mathbf{r})y\Theta(y)\Theta(R_c(\mathbf{r}) - R) \qquad (22)$$

This yields

$$F_x^{GGA}(s) = \frac{1}{9}\int_0^{z_c} dz z y_{sa}(s, z) \qquad (23)$$

where $z_c = 2k_f R_c$ and the spherical average (sa) of the positive part of y is

$$y_{sa}(s, z) = \frac{1}{4\pi}\int d\hat{\mathbf{R}}y\Theta(y) \qquad (24)$$

It is possible to obtain an analytic expression for y_{sa}, which is given explicitly in ref 64. However, the value of $z_c(s)$ and the final value of $F_x(s)$ must be found numerically.

To obtain the exact large s behavior, it can be shown that as s becomes large, the value of $z_c(s)$ becomes small, allowing the expression for $y_{sa}(s,z)$ to be replaced by a power series expansion. Retaining the dominant term we obtain $y_{sa} = (s^2z^2)/(27\sqrt{2})$. With this we can obtain the following analytic expressions for $z_c(s)$ and $F_x(s)$:

Table 1. Error in the Calculated Exchange-Only Interaction Energy ($Ex_{GGA} - E_{HF}$) at Several Separations Alongside the Full CCSD(T) Binding Energy[a]

separation (Å)	binding energy	HF	$Ex_{GGA} - E_{HF}$			
			PBE	revPBE	PW86	B86b
		H₂ Parallel (ref 59)				
3.55	0.04	0.067	-0.100^b	-0.014	-0.013	-0.026
4.05		0.020	-0.056^b	-0.032^b	-0.006	-0.019
4.55		0.008	-0.026^b	-0.021^b	-0.002	-0.010^b
		N₂ Crossed (ref 60)				
3.87	0.28	0.139	-0.078	0.274	0.086	0.100
4.37		0.028	-0.102^b	-0.026	0.014	-0.002
4.87		0.008	-0.057^b	-0.026^b	0.006	-0.011^b
		CO₂ Crossed (ref 61)				
3.25	0.62	1.13	-0.17	0.70	0.10	0.17
3.75		0.40	-0.29	0.01	-0.05	-0.10
4.25		0.20	-0.18	-0.09	-0.04	-0.08
		Ammonia (ref 58)				
3.16	3.17	-1.41	0.25	1.57	0.40	0.60
3.66		-1.66	-0.01	0.60	0.16	0.23
4.16		-1.15	-0.06	0.18	0.09	0.09
		Methane (ref 58)				
3.72	0.53	0.36	-0.12	0.46	0.07	0.13
4.22		0.07	-0.12^b	0.11	0.04	0.03
4.72		0.01	-0.09^b	-0.02^b	0.006	-0.02^b
		Ethene (ref 58)				
3.72	1.51	0.82	0.43	1.86	0.63	0.85
4.22		0.003	-0.02	0.62	0.22	0.27
4.72		-0.08	-0.12	0.12	0.08	0.05
		Benzene Parallel (ref 62)				
3.89	1.74	3.62	-0.63	0.98	-0.44	-0.04
4.39		1.56	-0.71	0.04	-0.48	-0.36
4.89		0.88	-0.60	-0.30	-0.41	-0.39
		Benzene T (ref 58)				
4.91	2.74	1.55	0.96	2.80	1.22	1.46
5.26		0.10	0.46	1.60	0.70	0.84
5.61		-0.30	0.17	0.84	0.41	0.46
		Pyrazine Parallel (ref 63)				
3.6	2.5	4.50	0.01	2.02	0.29	0.61
4.1		1.49	-0.35	-0.86	0.11	-1.37
4.6		0.67	-0.30	0.05	-0.04	-0.06

[a] Except in the CO₂ case where the result quoted is from MP2 calculations and the N₂ case where an experimental result is given. Energies are given in kcal/mol. [b] Values are cases of spurious exchange binding. Several dimers have attractive electrostatic interactions which dominate at large distances.

$$z_c(s) = \frac{3^{4/5}\sqrt{2}(5\pi)^{1/5}}{s^{2/5}} \simeq \frac{5.908}{s^{2/5}} \qquad (25)$$

and

$$F_x(s) = \frac{(5\pi/3)^{4/5}s^{2/5}}{3\sqrt{2}} \simeq 0.8863s^{2/5} \qquad (26)$$

Thus, we have shown that the exact large s dependence of the enhancement factor obtained from the PW86 integrals is $s^{2/5}$. The coefficient of 0.8863, obtained by exact evaluation the PW86 analytic expressions, is within 2% of the value implied by PW86's fitted parametrization (eq 11). Still, it is undoubtedly desirable to use rigorously correct results, wherever possible, and to that end, we use this result to refit the numerical results of the integration. We also correct the small s behavior which was originally constrained to recover $1 + (7/81)s^2$ (ref 65) but was later shown to correctly be $1 + (10/81)s^2$ (see, for example, refs 66 and 67). We keep the original form of the functional as given in eq 11, where a is set by the small s limit to be $10/81(= 0.1234)$, c is set by the large s behavior derived above to be 0.163 $(= 0.8863)$,[15] and the remaining parameter $b = 17.33$ is used to smoothly interpolate between the two limits and is obtained by fitting to the numerical solution for $F_x(s)$ for $0 \le s \le 10$. This constitutes a minor modification to the original PW86 form and does not have a substantial impact on the already good agreement PW86 exchange energies yielded with HF in the systems we have looked at. This can be seen in Figure 5 where the interaction energy of parallel N_2 molecules is shown for both the refit and the original PW86 functionals along with the HF results. The functional parameters are compared in Table 2. The large s asymptote derived here is probably the most robust feature of the analytic form for PW86, as it depends on the sum rule for the exchange hole but not on any of the other cutoffs that were necessarily imposed for evaluations for smaller reduced gradient. The situation may be different for some other functionals.[68] As s increases, it passes a threshold value above which the cutoff that enforces the negativity of the exchange hole is inoperative. We have calculated this threshold value numerically to be $s = 0.874$. The value of z_c, which enforces the second cutoff in eq 22, reduces with increasing s. At this threshold s value, z_c becomes low enough so that the function y_{sa}, described in eq 24, is restricted to a region where it is always positive, guaranteeing a negative exchange hole without any additional cutoffs.

Our analysis of the exchange-only interaction energy for a series of interacting molecules has shown that exchange functionals, which behave as $s^{2/5}$ at large s, offer a consistent improvement over other commonly used exchange functionals, with the PW86 functional being marginally better than Becke86b in the systems we have examined. The underlying origin of this $s^{2/5}$ dependence remains to be explained, although an enhancement factor that increases with s is necessary to mitigate the spurious LDA-like behavior identified by Harris (see discussion

Table 2. Parameters for the Original and Refit PW86 Functional As Described by Equation 11

	original PW86	refit PW86
a	0.0864	0.1234
b	14	17.33
c	0.2	0.163

following eq 5). We also note that, aside from B88, which is overly repulsive in these systems, the PW86 functional was the only functional tested that never yielded spurious exchange binding in the systems and the separations we examined. While B86b has a similar form to PW86, this indicates that the lower rate at which it increases with s, as can be seen in Figure 2, is not sufficiently large to rule out this type of spurious binding. Our analysis of the s-dependence of the exchange energies shows significant contributions are coming from highly inhomogeneous regions where s is $\gtrsim 10$, showing that the large s behavior is particularly important. We have derived the exact large s behavior of the integrals solved to obtain the PW86 functional and used this along with corrected small s behavior to refit the functional to the numerical results used to derive it. This is not a substantial change to the functional but allows it to more accurately reproduce the features of importance in the types of systems of interest.

Supporting Information Available: We have tabulated the errors incurred in the HF interaction energy by basis sets of increasing size relative to the CBS result for benzene in the T configuration and pyrazine in the parallel configuration. Full author list for reference 26. This material is available free of charge via the Internet at http://pubs.acs.org.

References

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

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

(3) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 5121–5129.

(4) Zhang, Y.; Vela, A.; Salahub, D. *Theor. Chem. Acc.* **2007**, *118*, 693–707.

(5) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748–8757.

(6) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

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

(8) Grimme, S.; Antony, J.; Schwabe, T.; Mueck-Lichtenfeld, C. *Org. Biomol. Chem.* **2007**, *5*, 741–758.

(9) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.

Exchange-Energy Density Functional Accuracy

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2761**

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

(11) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 014113.

(12) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. B: Condens. Matter* **2005**, *71*, 195119.

(13) Lin, I.-C.; Coutinho-Neto, M. D.; Felsenheimer, C.; von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U. *Phys. Rev. B: Condens. Matter* **2007**, *75*, 205131.

(14) Sun, Y. Y.; Kim, Y.-H.; Lee, K.; Zhang, S. B. *J. Chem. Phys.* **2008**, *129*, 154102.

(15) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.

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

(17) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104.

(18) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108.

(19) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.

(20) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.

(21) Johnson, E. R.; Becke, A. D. *Chem. Phys. Lett.* **2006**, *432*, 600–603.

(22) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 124108.

(23) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*−246401.

(24) Langreth, D. C.; Dion, M.; Rydberg, H.; Schröder, E.; Hyldgaard, P.; Lundqvist, B. I. *Int. J. Quantum Chem.* **2004**, *101*, 599.

(25) Thonhauser, T.; Cooper, V. R.; Li, S.; Puzder, A.; Hyldgaard, P.; Langreth, D. C. *Phys. Rev. B: Condens. Matter* **2007**, *76*, 125112.

(26) Langreth, D. C.; et al. *J. Phys.: Condens. Matter* **2009**, *21*, 084203.

(27) Kong, L.; Cooper, V. R.; Nijem, N.; Li, K.; Li, J.; Chabal, Y. J.; Langreth, D. C. *Phys. Rev. B: Condens. Matter* **2009**, *79*, 081407(R).

(28) Gulans, A.; Puska, M. J.; Nieminen, R. M. *Phys. Rev. B* **2009**, *79*, 201105(R).

(29) Vydrov, O. A.; Voorhis, T. V. *J. Chem. Phys.* **2009**, *130*, 104105.

(30) Aryasetiawan, F.; Miyake, T.; Terakura, K. *Phys. Rev. Lett.* **2002**, *88*, 166401.

(31) Fuchs, M.; Gonze, X. *Phys. Rev. B: Condens. Matter* **2002**, *65*, 235109.

(32) Kamiya, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 6010–6015.

(33) Miyake, T.; Aryasetiawan, F.; Kotani, T.; van Schilfgaarde, M.; Usuda, M.; Terakura, K. *Phys. Rev. B: Condens. Matter* **2002**, *66*, 245103.

(34) Aryasetiawan, F.; Miyake, T.; Terakura, K. *Phys. Rev. Lett.* **2003**, *90*, 189702.

(35) Fuchs, M.; Burke, K.; Niquet, Y.-M.; Gonze, X. *Phys. Rev. Lett.* **2003**, *90*, 189701.

(36) Furche, F.; Voorhis, T. *J. Chem. Phys.* **2005**, *122*, 164106.

(37) Ángyán, J. G.; Gerber, I. C.; Savin, A.; Toulouse, J. *Phys. Rev. A: At., Mol., Opt. Phys.* **2005**, *72*, 012510.

(38) Gerber, I. C.; Ángyán, J. G. *J. Chem. Phys.* **2007**, *126*, 044103.

(39) Sato, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 234114.

(40) Marini, A.; Garcia-Gonzalez, P.; Rubio, A. *Phys. Rev. Lett.* **2006**, *96*, 136404.

(41) Harl, J.; Kresse, G. *Phys. Rev. B: Condens. Matter* **2008**, *77*, 045136.

(42) Toulouse, J.; Gerber, I. C.; Jansen, G.; Savin, A.; Ángyán, J. G. *Phys. Rev. Lett.* **2009**, *102*, 096404.

(43) Lu, D.; Li, Y.; Rocca, D.; Galli, G. *Phys. Rev. Lett.* **2009**, *102*, 206411.

(44) Lacks, D. J.; Gordon, R. G. *Phys. Rev. A: At. Mol., Opt. Phys.* **1993**, *47*, 4681.

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

(46) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.

(47) Wang, Y.; Perdew, J. P. *Phys. Rev. B: Condens. Matter* **1986**, *33*, 8800(R).

(48) Becke, A. D. *J. Chem. Phys.* **1986**, *85*, 7184.

(49) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(50) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719.

(51) Harris, J. *Phys. Rev. B: Condens. Matter* **1985**, *31*, 1770.

(52) Zupan, A.; Burke, K.; Ernzerhof, M.; Perdew, J. P. *J. Chem. Phys.* **1997**, *106*, 10184.

(53) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524.

(54) Lieb, E. H.; Oxford, S. *Int. J. Quantum Chem.* **1981**, *19*, 427.

(55) Gonze, X.; et al. *Comput. Mater. Sci.* **2002**, *25*, 478.

(56) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(57) Gordon, M. S.; Schmidt, M. W. Advances in Electronic Structure Theory: Gamess a Decade Later. *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, The Netherlands, 2005; pp 1167−1189.

(58) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(59) Diep, P.; Johnson, J. K. *J. Chem. Phys.* **2000**, *112*−4465.

(60) Aquilanti, V.; Bartolomei, M.; Cappelletti, D.; Carmona-Novillo, E.; Pirani, F. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3891.

(61) Bock, S.; Bich, E.; Vogel, E. *Chem. Phys.* **2000**, *257*, 147.

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

(63) Sathyamurthy, N.; Dispersion, A. *J. Theor. Comput. Chem.* **2006**, *5*, 609–619.

(64) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B: Condens. Matter* **1996**, *54*, 16533.

(65) Sham, L. J. In *Computational Methods in Band Theory*; Marcus, P. M., Janak, J. P., Williams, A. R., Eds.; Plenum: New York, 1971; p 458.

(66) Kleinman, L.; Lee, S. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 4634.

(67) Engel, E.; Vosko, S. H. *Phys. Rev. B: Condens. Matter* **1990**, *42*, 4940.

(68) Ernzerhof, M.; Perdew, J. P. *J. Chem. Phys.* **1998**, *109*, 3313.

JCTC Journal of Chemical Theory and Computation

Beryllium Bonds, Do They Exist?

Manuel Yáñez,*,† Pablo Sanz,† Otilia Mó,† Ibon Alkorta,‡ and José Elguero‡

Departamento de Química, C-9, Universidad Autónoma de Madrid, Cantoblanco, E-28049-Madrid, Spain and Instituto de Química Médica, CSIC, Juan de la Cierva, 6, E-28006 Madrid, Spain

Abstract: The complexes between BeX_2 (X = H, F, Cl, OH) with different Lewis bases have been investigated through the use of B3LYP, MP2, and CCSD(T) approaches. This theoretical survey showed that these complexes are stabilized through the interaction between the Be atom and the basic center of the base, which are characterized by electron densities at the corresponding bond critical points larger than those found in conventional hydrogen bonds (HBs). Actually, all bonding indices indicate that, although these interactions that we named "beryllium bonds" are in general significantly stronger than HBs, they share many common features. Both interactions have a dominant electrostatic character but also some covalent contributions associated with a non-negligible electron transfer between the interacting subunits. This electron transfer, which in HBs takes place from the HB acceptor lone-pairs toward the σ_{YH}^* antibonding orbital of the HB donor, in beryllium bonds goes from the lone pairs of the Lewis base toward the empty p orbital of Be and the σ_{BeX}^* antibonding orbital. Accordingly, a significant distortion of the BeX_2 subunit, which in the complex becomes nonlinear, takes place. Concomitantly, a significant red-shifting of the X–Be–X antisymmetric stretching frequencies and a significant lengthening of the X–Be bonds occur. The presence of the beryllium bond results in a significant blue-shifting of the X–Be–X symmetric stretch.

Introduction

Noncovalent interactions play a very important role in chemistry, in particular as far as intermolecular interactions[1–3] and supramolecular chemistry[4–7] are concerned, where the molecular assemblies are usually held together through noncovalent and, quite often, weak interactions. Some of these interactions have special characteristics and, as a consequence, they constitute a subclass among the different kinds of intermolecular forces. This is the case of the conventional[8–11] and the nonconventional[12] hydrogen bonds as well as the so-called halogen bonds.[13–18] In these cases, the leading role of the interaction is played by a hydrogen or a halogen atom, which interacts with two moieties. For the particular case of the hydrogen bonds, these two moieties act as a Lewis base, the so-called HB acceptor, and as a

Lewis acid, which is the moiety usually called HB donor, respectively.

The nature of these interactions was not free of controversy, but it is generally accepted that the electrostatic component is dominant,[19–23] although it is not the only one responsible for the characteristics of these interactions. Actually, one of the necessary conditions to have an inter- or an intramolecular hydrogen bond is to have a hydrogen atom bonded to an electronegative element (X), so that the net positive charge on this hydrogen is usually rather high, favoring its electrostatic interaction with the basic site of the hydrogen bond acceptor, normally a heteroatom (Y) with a negative net charge.[9] Besides the electrostatic interaction between the hydrogen atom attached to X and Y, there are, unavoidably, the polarization effects and a non-negligible charge transfer from the lone-pairs of electrons of the HB acceptor toward the σ_{XH}^* antibonding orbital of the HB donor.[24] Recently, different analysis has clearly shown the role of this covalent contribution, which is reflected in the

---

* Corresponding author. Telephone: +34-91-4974953. Fax: +34-91-4975238. E-mail: manuel.yanez@uam.es.
† Universidad Autónoma de Madrid.
‡ Instituto de Química Médica.

value of delocalization indexes.[25] A very interesting analysis on the covalent/electrostatic nature of the hydrogen bond in terms of the interacting quantum atoms approach has been recently reported by Martín Pendás et al.[26] The aforementioned charge transfer process from the lone pairs of the HB acceptor toward the $\sigma_{XH}*$ antibonding orbital of the HB donor is necessarily reflected in a weakening of the X—H covalent bond, which accordingly lengthens, whereas its stretching frequency is red-shifted. This red shifting is precisely one of the experimental signatures of these interactions.[9,27−29]

These characteristics are not necessarily exclusive of hydrogen bonds, and they could be found whenever an element, not very electronegative, is covalently bonded to a more electronegative one and, at the same time, has low-lying empty orbitals which allow it to behave as a good electron acceptor. One of the elements which initially fulfill these conditions is beryllium. Due to its low electronegativity, it has been shown,[30] for instance, that BeH_2 may act as a proton-accepting molecule for dihydrogen-bonded systems. It is also well established, both from the experimental and theoretical viewpoints, that BeX_2 derivatives are good Lewis acids,[31,32] in particular, the $BeCl_2$ for which different complexes with nitrogen bases have been synthesized.[33] Also very recently, new complexes between $BeCl_2$ and dimethylsulfoxide have been isolated and characterized.[34] Actually $BeCl_2$ is a very strong Lewis acid, and somewhat unexpected, the $[BeCl_3]^-$ formed by the interaction of $BeCl_2$ with Cl^- is still capable of acting as a Lewis acid forming $[BeCl_3(L)]^-$ complexes, which have been recently reported to be stable.[35−39] The bonding in complexes involving $BeCl_2$ and $[BeCl_3]^-$ has been analyzed in detail by Frenking and co-workers[40] at different levels of theory, showing that the bond dissociation energy of Cl_2Be-NH_3 is larger than that of Cl_3B-NH_3, even though Be is less electronegative than B, and that the main difference in bonding comes from a larger electrostatic contribution in Be complexes than in boron ones.

The aim of this paper is to analyze the complexes between BeH_2 and other beryllium derivatives and different Lewis basis under a different perspective. Beryllium is a electropositive element, which in the limit could be regarded as a proton, which has a core of electrons; hence, the question we would like to answer is whether the aforementioned interactions involving Be derivatives can be considered a special kind of weak interactions, sharing some common characteristics with the conventional hydrogen and halogen bonds. In other words, we want to answer the question, is it possible to define "beryllium bonds" in a similar way as we define hydrogen bonds?

For this purpose, we have considered initially the interactions between BeX_2 (X = H, F, Cl, OH) derivatives with a set of small compounds, which are able to behave as Lewis bases and contain as basic sites first-, second- and third-row atoms, namely NH_3, H_2O, FH, PH_3, SH_2, ClH and BrH, (set A).

This initial set will be enlarged to a total of 13 reference bases (set B) by including, besides the bases of set A, some additional bases such as CO, HCN, LiCN, HN=CH_2, OCH_2, and CH_3NH_2. For this set B, we will use as a suitable example of the larger set of Be compounds investigated, the Cl_2Be derivative.

The possible existence of beryllium bonds may have some important implications in different fields. It is well established for instance that the association of beryllium fluoride to ADP inhibits protein action.[41−47] Be is also able to form half-sandwich complexes with interesting reactivity patterns.[48,49] Probably the most relevant aspect of our study concerns beryliosis or CBD (chronic beryllium disease) a very serious illness, often fatal, caused by Be metal and Be^{2+}. The mechanism is not clearly understood yet, but it involves the interaction of Be^{2+} with up to four water molecules.[50,51] In relation with CBD, beryllium water clusters have been studied theoretically[52] as well as the possibility that Be^{2+} could displace H^+ from strong hydrogen bonds.[53] Hence, the possible existence of beryllium bonds might offer alternative mechanisms to beryllium toxicity.

Computational Details

Standard MP2, CCSD(T), and density functional theory (DFT) calculations have been carried out by means of the Gaussian-03 suite of programs.[54] Among the different functionals nowadays available, we have chosen the B3LYP hybrid method,[55,56] which is likely the most frequently used because it yields quite reliable geometries and harmonic vibrational frequencies, so it would be interesting to assess its behavior as far as the description of "beryllium bonds" is concerned. For this purpose, the geometries of the complexes formed between Cl_2Be, taken as a benchmark case, and the Lewis basis included in set A were initially optimized at the B3LYP/6-311+G(d,p) and compared with those obtained when the MP2/6-311+G(d,p) approach was used. As shown in the Supporting Information, Table S1, the agreement between both set of values, as far as the length of the beryllium bond and the value of the ClBeCl angle, is very good, and only for second- and third-row basic atoms, the deviations between both sets of values are slightly larger. Hence, in what follows, we will use systematically B3LYP optimized geometries but refining the previous ones by using a larger 6-311+G(3df,2p) basis set expansion. The corresponding harmonic vibrational frequencies, which allowed us to characterize the stationary points as local minima of the potential energy surface and to estimate the thermal corrections to the enthalpy, and the Gibbs free energy were also obtained at the B3LYP/6-311+G(3df,2p) level of theory. The same level was used to evaluate the interaction energy for each complex. This interaction energy was defined as the dissociation energy of the "beryllium bond", i.e., as the negative of the enthalpy obtained by subtracting from the enthalpy of the complex the enthalpies of the two isolated monomers. The energies for the particular set of complexes formed by Cl_2Be, were also evaluated using the MP2 and CCSD(T) methods with an aug-cc-pVTZ basis set, for the sake of assessing the reliability of the DFT values. The CCSD(T) and DFT interaction energies were corrected with the basis set superposition error (BSSE) evaluated by means of the counterpoise method.[57]

The bonding of the systems under scrutiny was analyzed by means of three different approaches, namely the atoms

in molecules (AIM) theory,[58] the electron localization function (ELF) theory,[59,60] and the natural bond orbital (NBO) method.[24] In the framework of the AIM theory, we have evaluated the electron density at the different bond critical points (BCP), since the existence of a BCP in the region between the Be atom and the basic site of the Lewis base will permit not only to establish the existence of a "beryllium bond" but also to quantitatively estimate its strength. Besides, the energy density at the BCP provides useful clues about the covalent nature of the interaction.

By means of the ELF theory,[59,60] it is possible to partition the molecular space in basins of two types, polysynaptic basins (generally disynaptic), with the participation of two (or more) atomic valence shells, and monosynaptic ones, which correspond to electron lone-pairs,[61] because the ELF function becomes large in regions of space where electron pairs, either bonding or lone pairs, are localized. ELF grids and basin integrations have been evaluated with the TopMod package.[62] For the three-dimensional plots, an ELF value of 0.8 has been used.

The NBO approach[24] describes the bonding in terms of localized hybrids and lone pairs obtained as local block eigenvectors of the one-particle density matrix. Besides, a second-order perturbation analysis of the Fock matrix permits to quantify the interactions between occupied MOs of the Lewis base and unoccupied MOs of the Lewis acid, in a typical electron transfer process from the former toward the latter. The Wiberg bond orders (BO)[63] were also evaluated as defined in the NBO-5G program package.[64]

Results and Discussion

Full optimized geometries of the complexes formed by the association of BeX_2 (X = H, F, Cl, OH) derivatives and the compounds of set A are collected in the Supporting Information, Table S2. The interaction enthalpies obtained at the B3LYP/6-311+G(3df,2p) level are summarized in Table 1.

The first conspicuous fact is that the length of the "beryllium bond", considered as the internuclear distance between Be and the basic site (Y = C, N, O, F, P, S, Cl, Br) of the Lewis base, is significantly smaller than the sum of the van der Waals radii of the two atoms (See Table 1), which clearly indicates the existence of a significant interaction between the two monomers forming the complex. The deformation undergone by the BeX_2 compound, which in the complex is not linear anymore, is a further indication of the strength of these interactions (See Table 1). The second important feature is that the interaction enthalpies are very large for most of the bases, and they are only small when the basic site is Cl or Br. More importantly, these interaction enthalpies are in some cases four or five times larger than the energies involved in typical hydrogen bonds. It must be reminded here, for instance, that the interaction energy in the water dimer is about 20 kJ mol^{-1},[65−67] whereas the enthalpy of some of the beryllium bonds investigated here are greater than 100 kJ mol^{-1}.

Statistically, it is possible to analyze the ΔH_{int} values of Table 1 with regard to NH_3 as reference base (R2 = 0.991), in terms of the presence/absence Free-Wilson matrix[68,69] which is given in the Supporting Information. The values of

Table 1. B3LYP/6-311+G(3df,2p) Optimized Be−Y Distances ($R_{Be−Y}$ in Å), XBeX Angles (in degrees), and Dissociation Enthalpies (ΔH_{int}, kJ mol^{-1}) of Different Complexes Involving BeX_2 (X = H, F, Cl, OH) Derivatives with Lewis bases of the Set A

complex	$R_{Be−Y}$	XBeX angle	ΔH_{int}
H_2Be-NH_3	1.771	138.6	86
H_2Be-OH_2	1.666	143.6	69
H_2Be-FH	1.798	152.1	25
H_2Be-PH_3	2.285	145.6	24
H_2Be-SH_2	2.251	147.6	24
$H_2Be-ClH$	2.378	157.0	6
$H_2Be-BrH$	2.611	160.0	4
F_2Be-NH_3	1.776	137.6	103
F_2Be-OH_2	1.690	142.1	81
F_2Be-FH	1.802	150.2	35
F_2Be-PH_3	2.342	144.9	30
F_2Be-SH_2	2.306	146.3	31
$F_2Be-ClH$	2.377	154.3	15
$F_2Be-BrH$	2.603	156.8	11
Cl_2Be-NH_3	1.745	134.7	106
Cl_2Be-OH_2	1.653	139.2	81
Cl_2Be-FH	1.748	145.7	28
Cl_2Be-PH_3	2.264	139.3	29
Cl_2Be-SH_2	2.244	140.8	27
$Cl_2Be-ClH$	2.338	148.2	4
$Cl_2Be-BrH$	2.602	151.4	1
$(OH)_2Be-NH_3$	1.770	143.4	84
$(OH)_2Be-OH_2$	1.700	148.3	64
$(OH)_2Be-FH$	1.797	155.8	32
$(OH)_2Be-PH_3$	2.336	151.4	14
$(OH)_2Be-SH_2$	2.306	152.8	16
$(OH)_2Be-ClH$	2.470	162.6	5
$(OH)_2Be-BrH$	2.802	168.0	14

this statistical analysis (H_2Be 91.3, F_2Be 101.0, Cl_2Be 96.7, $(OH)_2Be$ 90.0, OH_2 −21.0, FH −64.8, PH_3 −70.5, SH_2 −70.3, ClH −87.3 and BrH −87.3 kJ mol^{-1}) are very close to those in Table 1, which is an indication of the goodness of this analysis. All the bases have negative coefficients because NH_3 is giving the strong interaction enthalpies; then, the order, as far as interaction energies are concerned, is NH_3 > OH_2 > FH > SH_2 > PH_3 > ClH = BrH. Concerning the beryllium derivatives, the order is F_2Be > Cl_2Be > H_2Be > $(OH)_2Be$.

In this respect, it is worth to compare the stability of the beryllium bonds with respect to the hydrogen bonds, in which the latter can be also formed between the two interacting systems. To analyze this question, the subset of complexes formed by $Be(OH)_2$ is very appropriate since this compound can act as either a HB donor or acceptor. In Figure 1, we present the structures of the two kind of possible complexes, those stabilized through beryllium and hydrogen bonds as well as the calculated interaction energies. It can be noted that $(OH)_2Be$ behaves as a HB donor with respect to NH_3, OH_2, PH_3 and SH_2, which are stronger HB acceptors, whereas it behaves as a HB acceptor when interacting with FH, ClH and BrH, which are stronger HB donors. Only for the case of water, both complexes were found to be almost equally stable. The important finding to be underlined here is that beryllium bonds are predicted to be, by far, more stabilizing than the HBs, and only for FH, ClH and BrH, both interactions are energetically similar. Nevertheless, as we shall show in forthcoming sections, even in these three cases,
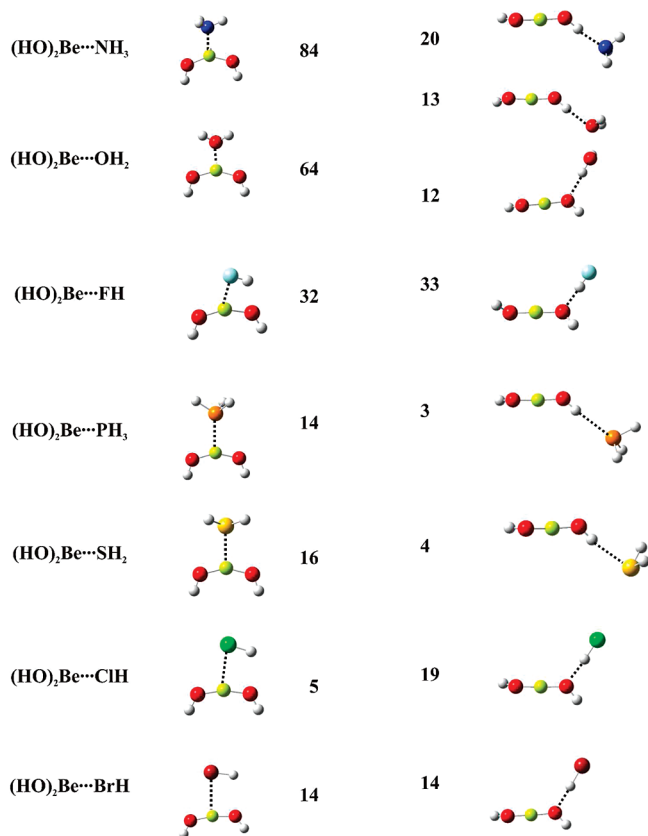
Figure 1. Optimized structures of the two kind of complexes formed by the interaction of Be(OH)$_2$ with the Lewis bases included in set A. The left-hand side column corresponds to the complexes stabilized through a beryllium bond, while those in the right-hand side column are hydrogen bonded complexes. The interaction enthalpies between both subunits are given by each structure in kJ mol^{-1}.

the beryllium bond is predicted to be stronger than the hydrogen bond, when evaluated at the CCSD(T) level.

Bonding

The origin of the strength of these intermolecular interactions, that we have named beryllium bonds, requires an analysis of their intimate nature. For this purpose, we will examine, as a suitable benchmark case, the complexes formed between BeCl$_2$ and the bases included in set B. The B3LYP/6-311+G(3df,2p) optimized geometries are shown in Figure 2. The interaction enthalpies, obtained at different levels of theory, have been summarized in Table 2.

The first important point to be emphasized is that all Be complexes investigated share the same bonding characteristics, although only those the BeCl$_2$ complexes will be discussed in detail. To better illustrate this similarity we have included in the Supporting Information, Figure S1, a comparison between the molecular graphs and the ELFs of BeCl$_2$ and BeH$_2$ complexes.

A topological analysis of the electron density of the systems under scrutiny shows the existence of a BCP associated with the beryllium bond. The electron density associated to these BCPs is typical of the interactions between closed shell systems but larger than those found for conventional HBs. It is also worth noting that the energy
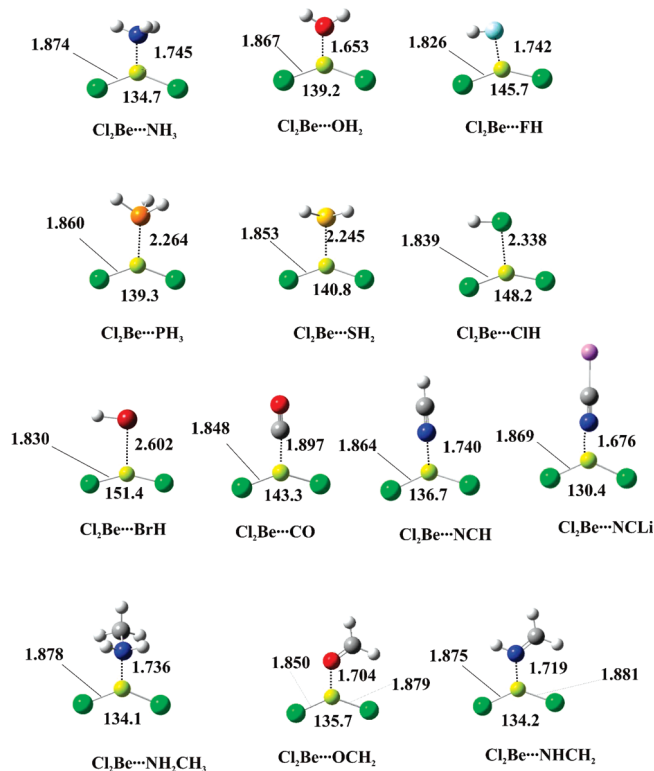
Figure 2. Optimized geometries for the complexes between BeCl$_2$ and the Lewis bases included in set B. Bond lengths are in Å and bond angles in degrees.

density, although small in absolute value, is negative in most cases (See Table 3), pointing out a non-negligible covalent contribution to these interactions. Note that only for very strong HBs, the energy density is slightly negative.

This AIM description is in harmony with the NBO picture. As shown in Table 3, the Wiberg BO of the beryllium bonds is quite significant in some cases (close or greater than 0.3). It must be recalled here, that Wiberg BOs of polar covalent bonds are generally small. Actually, the Wiberg BO for the Cl—Be polar covalent bond in the isolated BeCl$_2$ molecule is 0.44, not much greater than some of the values calculated for the beryllium bonds. No doubt, one important energetic contributor to the beryllium bonds is of electrostatic origin. As a matter of fact, the net natural charge of Be in the complexes investigated is systematically around +0.5, whereas the basic site of the different bases bear a negative net natural charge, which varies from −0.87 in CH$_3$NH$_2$ to −0.16 in HBr. However, there is not a direct correlation between the net charge of the basic site and the interaction energy, and, even in the case of CO complexes where the carbon atom bears a positive net charge (+0.39), still the interaction energy is of the order of 30 kJ mol^{-1}. Hence, other factors besides the electrostatic interactions should contribute to the stability of the beryllium bond. A NBO second-order perturbation analysis actually shows the existence of a quite large electron transfer from the lone-pairs of the Lewis base toward the empty p orbital in Be, and the σ_{ClBe}^* antibonding molecular orbital (see Table 3). As it is clearly reflected in Table 3, these second-order orbital interactions decrease as the electron donor ability of the base decrease, and therefore, they are smaller in FH complexes

Table 2. Energetics of Complexes Between Cl_2Be and Different Lewis Bases[a]

complex	ΔH_0(adi.) B3LYP	ΔH_0(adi.) MP2	ΔH_0(adi.) CCSD(T)	ΔH_{relax} CCSD(T)	ΔH_0 (vert.) CCSD(T)
Cl_2Be-NH_3	106 (104)	126	125 (118)	54	179
Cl_2Be-OH_2	81 (78)	97	96 (90)	47	143
Cl_2Be-FH	28 (24)	39	39 (32)	32	71
Cl_2Be-PH_3	29 (27)	49	44 (39)	48	92
Cl_2Be-SH_2	27 (25)	47	43 (38)	38	81
$Cl_2Be-ClH$	4 (1)	20	17 (13)	24	41
$Cl_2Be-BrH$	1 (0)	22	19 (11)	19	38
$Cl_2Be-HCN$	57 (55)	75	71 (64)	48	119
$Cl_2Be-OCH_2$	62 (59)	82	82 (75)	53	135
$Cl_2Be-CO-a$	0	0	0		
$Cl_2Be-CO-b$	18 (16)	32	26 (20)	33	59
$Cl_2Be-LiCN$	130 (127)	149	147 (139)	67	214
$Cl_2Be-NHCH_2$	115 (113)	139	136 (128)	58	194
$Cl_2Be-MeNH_2$	117 (115)	147	145 (137)	58	203

[a] ΔH_0 (adi.), ΔH_{relax}, ΔH_0 (vert.) stand for the adiabatic dissociation energy, the relaxation energy of the products and the vertical dissociation energy. The values within parentheses correspond to the dissociation energies after the BSSE correction is included. All values are in kJ mol^{-1}.

Table 3. Electron Density (ρ_{BCP}, au) and Energy Density (au) at the Be$\cdots$Y BCP, Electron Population at the V(Be,Y) and V(Cl,Be) Disynaptic Basins, Wiberg BO Associated with the Be$\cdots$Y Interaction, and the Second-Order Interaction Energies (kJ mol^{-1}) between the Lone-Pairs of the Bases and the Empty p Orbital of Be and the $\sigma_{ClBe}{}^*$ Antibonding Orbital for the Complexes Formed between $BeCl_2$ and the Lewis Bases Included in Set B

complex	ρ_{BCP} (energy density)	V(Be,Y) (V(Cl,Be)) basin populations	BO	$LP_{base} \rightarrow p_{Be}$	$LP_{base} \rightarrow \sigma_{ClBe}{}^*$
Cl_2Be-NH_3	0.069 (−0.009)	2.08 (1.81)	0.23	87	58
Cl_2Be-OH_2	0.064 (+0.004)	2.38 (1.81)	0.10	60	56
Cl_2Be-FH	0.042 (+0.003)	1.15 (1.86)	0.08	34	38
Cl_2Be-PH_3	0.042 (−0.011)	2.06 (1.84)	0.24	Be−P bond[b]	
Cl_2Be-SH_2	0.038 (−0.008)	2.01 (1.85)	0.20	Be−S bond[c]	
$Cl_2Be-ClH$	0.025 (−0.004)	−[a] (1.90)	0.12	32	33
$Cl_2Be-BrH$	0.020 (−0.004)	−[a] (1.92)	0.11	21	24
$Cl_2Be-HCN$	0.057 (0.000)	3.08 (1.84)	0.12	71	56
$Cl_2Be-OCH_2$	0.056 (+0.003)	2.48 (1.84)	0.11	85	80
Cl_2Be-CO	0.046 (−0.005)	2.59 (1.87)	0.19	Be−C bond[d]	
$Cl_2Be-NCLi$	0.071 (−0.006)	3.08 (1.82)	0.14	105	58
$Cl_2Be-NHCH_2$	0.069 (−0.008)	2.59 (1.83)	0.13	75	75
$Cl_2Be-MeNH_2$	0.069 (−0.010)	2.15 (1.81)	0.33	Be−N bond[e]	

[a] No V(Be,Cl) disynaptic basin has been located. [b] 6% Be (31% s + 69% p) + 94% P (32% s + 68% p). [c] 5% Be (29% s + 71% p) + 95% S (15% s + 85% p). [d] 4% Be (35% s + 65% p) + 96% C (67% s + 33% p). [e] 9% Be (28% s + 72% p) + 91% N (35% s + 65% p).

than in NH_3 ones and they are very large for very basic compounds, such as LiCN, $NH=CH_2$ and $MeNH_2$. In some cases, the NBO approach actually detects the formation of a covalent bond between the Be and the basic site of the base, but this is likely a consequence of the localization scheme used by this method. In these cases, as indicated in Table 3, the situation is not essentially difference for the other ones, in the sense that the bond located implies a rather small contribution from the hybrid on Be, which is compatible with the description above in terms of an electron transfer from the lone-pairs of the base toward the empty orbitals of Be. This bonding similarity is also mirrored in the analogy of the ELF for the ammonia and the phosphine complexes (see Figure 3). This electron transfer is obviously reflected in the electron population of the initially empty p orbital of Be, which in the complexes investigated is typically around 0.06 *e*. This participation of the p orbital in the complex is actually responsible for the deformation undergone by the $BeCl_2$ subunit, which accordingly is not linear any more. Similarly the electron transfer toward the $\sigma_{ClBe}{}^*$ antibonding molecular orbital is reflected in a significant lengthening of

the Cl−Be with respect to the isolated molecule (see Figure 2) as well as in a decrease of the electron density at the Be−Cl BCP (see Table 4) and of the population of the V(Be,Cl) disynaptic basin (See Figure 3).

The formation of the complex also has a strong effect in the frequencies of the $BeCl_2$ subunit. Both in the isolated molecule and the complexes, the Be−Cl stretching modes appear as symmetric and antisymmetric combinations. Consistently, with the weakening of the Be−Cl bonds upon complexation, the antisymmetric combination appears significantly shifted to the red (See Table 4). The larger the shifting, the stronger the interaction is between the lone pair of the base and the $\sigma_{ClBe}{}^*$ orbital. However, the symmetric combination is significantly blue-shifted. This is so because in the complex the $BeCl_2$ subunit is bent, and therefore the symmetric stretch of the Be−Cl bonds is unavoidably coupled with the stretching of the Be−Y bond. In other words, the symmetric elongation of the two Be−Cl bonds leads to a compression of the Be−Y and vice versa. Hence, the stronger the beryllium bond, the larger the blue shifting of the BeCl symmetric stretch
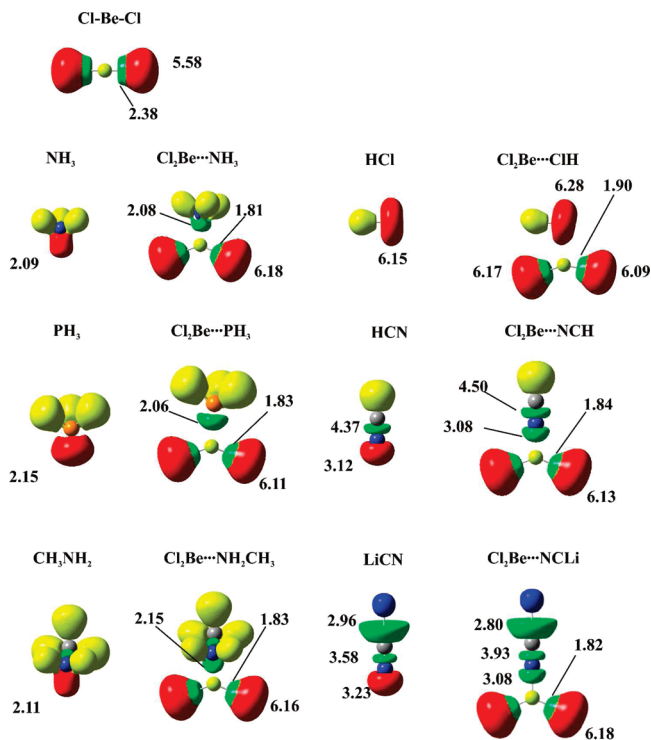
Figure 3. Three-dimensional representations of ELF isosurfaces with ELF = 0.80 for Cl_2Be-containing complexes. Yellow lobes correspond to disynaptic basins involving H atoms, red lobes correspond to monosynaptic lone-pair basins, green lobes correspond to disynaptic basins between two bonded atoms, and blue lobes correspond to monosynaptic atomic cores. The population of the different basins is given in e^-.

Table 4. Be–Cl Stretching Frequencies (cm^{-1}) and Electron Density at the Be–Cl BCP (ρ_{BCP}, au)

	ν_{Cl-Be} (sym. stret.)	ν_{Cl-Be} (asym. stret.)	ρ_{BCP}
isolated $BeCl_2$	339	1 131	0.095
complex			
NH_3	719	928	0.078
H_2O	630	954	0.080
FH	507	1 008	0.089
PH_3	539	923	0.082
SH_2	523	945	0.083
ClH	461	1 003	0.086
BrH	440	1 025	0.088
HCN	635	913	0.080
CH_2O	623	920	0.080
CO	547	954	0.083
LiCN	752	850	0.075
$NH=CH_2$	726	885	0.076
$MeNH_2$	726	897	0.077

becomes; as it is nicely illustrated by the rather good exponential correlation between the blue shifting and the electron density at the beryllium bond BCP (see Figure 4).

The existence of a beryllium bond is also ratified by the ELF analysis, which shows the existence of a disynaptic V(Be,Y) basin in most of the complexes investigated. As a matter of fact, only in those cases where the compound interacting with the $BeCl_2$ molecule is a poor Lewis base, like HCl, these disynaptic basins are not found, and only a polarization of the basic site lone-pairs is observed (See
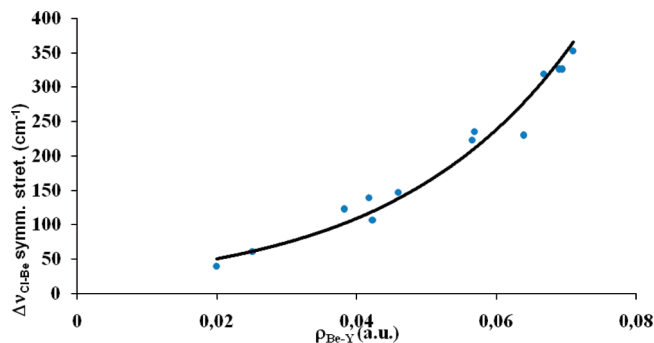
Figure 4. Exponential correlation between the blue shifting of the Cl–Be–Cl symmetric stretch and the electron density at the beryllium bond BCP. This correlation fulfils the equation: $\Delta\nu_{Cl-Be}$ 2.82 exp (39.04 ρ_{Be-Y}) (R^2 = 0.966).

Figure 3). It is also evident from Figure 3 that lone pairs of the basic site undergo a dramatic volume contraction on going from the isolated base to the complex. For example, upon the interaction between Cl_2Be and NH_3, the volume of the N lone pair decreases from 92 to 34 au, even though its population hardly changes. This is a direct consequence of the strong interaction between the lone pair and the empty beryllium orbitals (actually in the complex the lone pair becomes a V(N,Be) disynaptic basin), which leads to very short distances between the basic site of the Lewis base and the Be atom and, therefore, to a dramatic volume contraction of the base lone pair.

The significant geometry distortions triggered by the complexation process must have a large effect on the calculated enthalpy[40] of the beryllium bond because this is defined with respect to the isolated and, therefore, unperturbed monomers. A more realistic measure of the strength of the beryllium bonds would be obtained by calculating the vertical dissociation energy of the complex, i.e., the energy required to separate the interacting subunits at infinite distance, but keeping frozen the structure they have in the complex. This will be equivalent to adding to the adiabatic dissociation enthalpy the enthalpy involved in the relaxation of the two subunits from their structure within the complex to their equilibrium structure. As shown in Table 2, this relaxation enthalpy (ΔH_{relax}) is quite large and, for the weaker complexes, almost equal or even slightly larger than the adiabatic dissociation energy itself. It is important to note that the larger the relaxation or deformation energy the greater the stability of the complex, a fact which is nicely reflected in the good correlation between the adiabatic and the vertical dissociation energies (See Figure 5). It is also worth noting the very good agreement of our calculated values for the Cl_2Be-NH_3 complex and those reported previously in the literature.[40]

Finally, it is important to notice that the B3LYP adiabatic dissociation energies, ΔH_0(adi.), underestimate the values obtained at the CCSD(T) level. A similar result was reported by Frenking and co-workers for the particular case of the Cl_2Be-NH_3 complex when the BP86 functional was used.[40] It is worth noting that the agreement between B3LYP and CCSD(T) results improve slightly when including the BSSE correction because, at the CCSD(T) level this error is 6.7 kJ
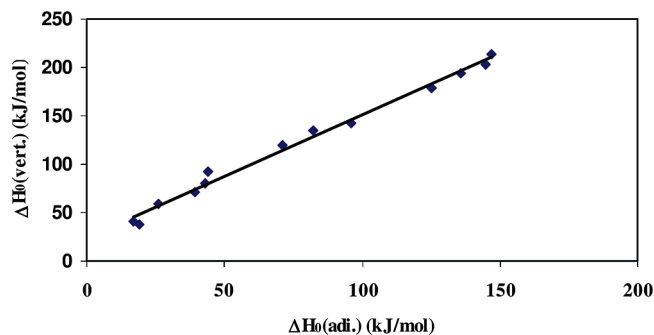
Beryllium Bonds

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2769**

Figure 5. Linear correlation between the vertical and the adiabatic dissociation energies of the complexes formed by $BeCl_2$ with the Lewis bases of set B. This correlation fulfils the equation ΔH_0 (vert.) = 1.265 ΔH_0 (adi.) + 24.32 kJ mol^{-1} (R^2 = 0.991).

mol^{-1}, in average, whereas for the B3LYP calculations is about three times smaller (2.4 kJ mol^{-1}, in average).

In order to establish whether a high level of electron correlation is required to adequately describe these bonds, we have also compared the CCSD(T) and B3LYP values with those obtained at the MP2 level. In general, the agreement between CCSD(T) and MP2 values is very good, and hence, one may rely on the lower level of theory to get reasonably good estimates for the energies of beryllium bonds. As a matter of fact, as shown in the Supporting Information, Figure S2, the correlation between both sets of values is very good. The correlation between B3LYP and CCSD(T) estimates is slightly poorer (see Supporting Information, Figure S3) but still very good, which means that the B3LYP method correctly reproduces the stability trends obtained with the much more expensive CCSD(T) approach. Finally, it should be mentioned that, in spite of the differences in the energetics obtained with B3LYP and MP2 (or CCSD(T)) methods, the bonding picture obtained from B3LYP and MP2 (or CCSD) densities is essentially the same. The Wiberg BOs and the electron densities at the BCP vary by less than 0.01 and 0.001 au, respectively. Similarly the populations of the ELF basins change inappreciably as illustrated in the Supporting Information, Figure S4, where the ELFs obtained at the MP2 level for four different complexes, taken as suitable examples, are compared with the B3LYP ones.

Conclusions

The complexes between BeX_2 (X = H, F, Cl, OH) with different Lewis bases are stabilized through the interaction between the Be atom and the basic center of the base. These interactions, which we have called beryllium bonds, are characterized by the existence of a BCP with electron densities similar to other typical closed-shell interactions but with energy densities, which in some cases are slightly negative, indicating the presence of non-negligible covalent contributions. All bonding indices actually indicate that, although beryllium bonds are in general significantly stronger than HBs, they share many common features.

Although both interactions have a dominant electrostatic character they have also some covalent character associated with a non-negligible electron transfer between the interacting subunits. In HBs, this electron transfer takes place from the HB acceptor lone-pairs toward the σ_{YH}* antibonding orbital of the HB donor, whereas in beryllium bonds they go from the lone pairs of the Lewis base toward the empty p orbital of the Be and the σ_{BeX}* antibonding orbital. This electron transfer is reflected in a significant distortion of the BeX_2 subunit, which in the complex becomes nonlinear and in a significant red-shifting of the X-Be-X antisymmetric stretching frequencies. The presence of the beryllium bond results in a significant blue-shifting of the X−Be−X symmetric stretch. Hence, also similarly to HBs whose experimental signature is the red-shifting of the XH stretching frequency of the HB donor,[27−29] in beryllium bonds, this signature should be the red-shifting of the X−Be−X antisymmetric stretch and the blue-shifting of the symmetric one.

Supporting Information Available: Comparison between B3LYP and MP2 optimized geometries, B3LYP/6-311+G(3df,2p) optimized geometries of $X_2Be\cdots YH_n$ (X = H, F, OH, Cl; Y = N, O, F, P, S, Cl, Br) complexes, presence−absence matrix corresponding to Table 1 data, molecular graphs and ELFs of H_2Be and Cl_2Be complexes, correlations between dissociation enthalpies calculated at the CCSD(T), MP2, and B3LYP levels, and comparison between the ELFs obtained using B3LYP and MP2 electron densities. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Buckingham, A. D.; Fowler, P. W.; Hutson, J. M. *Chem. Rev.* **1988**, *88*, 963–988.

(2) Chalasinski, G.; Szczesniak, M. M. *Chem. Rev.* **2000**, *100*, 4227–4252.

(3) Wormer, P. E. S.; van der Avoird, A. *Chem. Rev.* **2000**, *100*, 4109–4143.

(4) Philp, D.; Stoddart, J. F. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1155–1196.

(5) Rudkevich, D. M. *Angew. Chem., Int. Ed.* **2004**, *43*, 558–571.

(6) Saalfrank, R. W.; Maid, H.; Scheurer, A. *Angew. Chem., Int. Ed.* **2008**, *47*, 8794–8824.

(7) Llanes-Pallas, A.; Palma, C. A.; Piot, L.; Belbakra, A.; Listorti, A.; Prato, M.; Samori, P.; Armaroli, N.; Bonifazi, D. *J. Am. Chem. Soc.* **2009**, *131*, 509–520.

(8) Pimentel, G. C.; McClellan, A. L. *Hydrogen Bond*; W.H. Freeman: San Francisco, CA, 1960.

(9) Jeffrey, G. A. *An Introduction to Hydrogen Bonding*; Oxford University Press: New York, 1997.

(10) Jeffrey, G. A. *Crystal. Rev.* **2003**, *9*, 135–176.

(11) Grabowski, S. J. *Hydrogen Bonding -New Insights* Springer: Amsterdam, The Netherlands, 2006; Vol. 3.

(12) Alkorta, I.; Rozas, I.; Elguero, J. *Chem. Soc. Rev.* **1998**, *27*, 163–170.

(13) Zou, J. W.; Jiang, Y. J.; Guo, M.; Hu, G. X.; Zhang, B.; Liu, H. C.; Yu, Q. S. *Chem.—Eur. J.* **2005**, *11*, 740–751.

(14) Aakeroy, C. B.; Fasulo, M.; Schultheiss, N.; Desper, J.; Moore, C. *J. Am. Chem. Soc.* **2007**, *129*, 13772.

(15) Riley, K. E.; Hobza, P. *J. Chem. Theor. Comp.* **2008**, *4*, 232–242.

(16) Alkorta, I.; Blanco, F.; Elguero, *J. Struct. Chem.* **2009**, *20*, 63–71.

(17) Bernal-Uruchurtu, M. I.; Hernandez-Lamoneda, R.; Janda, K. C. *J. Phys. Chem. A* **2009**, *113*, 5496–5505.

(18) Metrangolo, P.; Meyer, F.; Pilati, T.; Resnati, G.; Terraneo, G. *Angew. Chem., Int. Ed.* **2008**, *47*, 6114–6127.

(19) Morokuma, K. *Acc. Chem. Res.* **1977**, *10*, 294–300.

(20) Minyaev, R. M.; Orlova, G. V. *J. Struct. Chem.* **1985**, *26*, 157–164.

(21) Jorgensen, W. L.; Severance, D. L. *J. Am. Chem. Soc.* **1991**, *113*, 209–216.

(22) Chen, W.; Gordon, M. S. *J. Phys. Chem.* **1996**, *100*, 14316–14328.

(23) Uchimaru, T.; Korchowiec, J.; Tsuzuki, S.; Matsumura, K.; Kawahara, S. *Chem. Phys. Lett.* **2000**, *318*, 203–209.

(24) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.

(25) Poater, J.; Fradera, X.; Sola, M.; Duran, M.; Simon, S. *Chem. Phys. Lett.* **2003**, *369*, 248–255.

(26) Martín Pendás, A.; Blanco, M. A.; Francisco, E. *J. Chem. Phys.* **2006**, *125*.

(27) Badger, R. M.; Bauer, S. H. *J. Chem. Phys.* **1937**, *5*, 839–851.

(28) Huggins, C. M.; Pimentel, G. C. *J. Phys. Chem.* **1956**, *60*, 1615–1619.

(29) Ohno, K.; Okimura, M.; Akai, N.; Katsumoto, Y. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3005–3014.

(30) Grabowski, S. J. *J. Mol. Struct.* **2000**, *553*, 151–156.

(31) Holleman, F.; Wiberg, N.; Wiber, E. Inorganic Chemistry; Academic Press: Cambridge, 2001.

(32) Housecroft, C.; Sharpe, A. G. Inorganic Chemistry, 3rd ed.; Prentice Hall: New York, 2008.

(33) Dressel, M. P.; Nogai, S.; Berger, R. J. F.; Schmidbaur, H. *Z. Naturforsch.(B)* **2003**, *58*, 173–182.

(34) Neumuller, B.; Petz, W.; Dehnicke, K. *Z. Anorg. Allg. Chem.* **2008**, *634*, 662–668.

(35) Neumuller, B.; Weller, F.; Dehnicke, K. *Z. Anorg. Allg. Chem.* **2003**, *629*, 2195–2199.

(36) Neumuller, B.; Dehnicke, K. *Z. Anorg. Allg. Chem.* **2003**, *629*, 2529–2534.

(37) Neumuller, B.; Dehnicke, K. *Z. Anorg. Allg. Chem.* **2004**, *630*, 347–349.

(38) Neumuller, B.; Dehnicke, K. *Z. Anorg. Allg. Chem.* **2004**, *630*, 369–376.

(39) Neumuller, B.; Dehnicke, K. *Z. Anorg. Allg. Chem.* **2004**, *630*, 799–805.

(40) Metz, S.; Holthausen, M. C.; Frenking, G. *Z. Anorg. Allg. Chem.* **2006**, *632*, 814–818.

(41) Chen, X. R.; Grammer, J.; Lawson, J. D.; Cooke, R.; Pate, E.; Yount, R. G. *Biochemistry* **2002**, *41*, 2609–2620.

(42) Shibuya, H.; Kondo, K.; Kimura, N.; Maruta, S. *J. Biochem.* **2002**, *132*, 573–579.

(43) Danko, S.; Yamasaki, K.; Daiho, T.; Suzuki, H. *J. Biol. Chem.* **2004**, *279*, 14991–14998.

(44) Sauna, Z. E.; Nandigama, K.; Ambudkar, S. V. *J. Biol. Chem.* **2006**, *281*, 26501–26511.

(45) Tiago, T.; Simao, S.; Aureliano, M.; Martin-Romero, F. J.; Gutierrez-Merino, C. *Biochemistry* **2006**, *45*, 3794–3804.

(46) Orban, J.; Lorinczy, D.; Hild, G.; Nyitrai, M. *Biochemistry* **2008**, *47*, 4530–4534.

(47) Liu, F.; Putnam, A.; Jankowsky, E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20209–20214.

(48) Barbaro, P.; Cecconi, F.; Dakternieks, D.; Dominguez, S.; Duthie, A.; Ghilardi, C. A.; Midollini, S.; Orlandini, A.; Vacca, A. *Inorg. Chem.* **2001**, *40*, 2725–2729.

(49) Conejo, M. D.; Fernandez, R.; Carmona, E.; Andersen, R. A.; Gutierrez-Puebla, E.; Monge, M. A. *Chem.—Eur. J.* **2003**, *9*, 4462–4471.

(50) Schmidbaur, H. *Coord. Chem. Rev.* **2001**, *215*, 223–241.

(51) Scott, B. L.; McCleskey, T. M.; Chaudhary, A.; Hong-Geller, E.; Gnanakaran, S. *Chem. Commun.* **2008**, 2837–2847.

(52) Asthagiri, D.; Pratt, L. R. *Chem. Phys. Lett.* **2003**, *371*, 613–619.

(53) McCleskey, T. M.; Ehler, D. S.; Keizer, T. S.; Asthagiri, D. N.; Pratt, L. R.; Michalczyk, R.; Scott, B. L. *Angew. Chem., Int. Ed.* **2007**, *46*, 2669–2671.

(54) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; J. A. Montgomery, J.; 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.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, J.; Cammi, R.; Pomelli, C.; Ochterski, J.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A.; *Gaussian03*, Revision E.01; Gaussian, Inc.: Wallingford, CT, 2003.

(55) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–7.

(56) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785–9.

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

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

(59) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397–403.

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

(61) Silvi, B. *Phys. Chem. Chem. Phys.* **2004**, *6*, 256–260.

(62) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem.* **1999**, *23*, 597–604, TopMoD package, Université Pierre et Marie Curie, 1997.

(63) Wiberg, K. B. *Tetrahedron* **1968**, *24*, 1083–1096.

(64) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F.; *Natural Bond Orbital NBO 5.0*. Theoretical Chemistry Institute, University of Wisconsin: Madison, WI. http://www.chem.wisc.edu/~nbo5 ed; 2004.

(65) Feyereisen, M. W.; Feller, D.; Dixon, D. A. *J. Phys. Chem.* **1996**, *100*, 2993–2997.

(66) Fiadzomor, P. A. Y.; Keen, A. M.; Grant, R. B.; Orr-Ewing, A. J. *Chem. Phys. Lett.* **2008**, *462*, 188–191.

(67) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer, H. F.; Quack, M. *J. Chem. Phys.* **2002**, *116*, 690–701.

(68) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, Vol. 1; Heller, S. R., Ed.; Oxford University Press: New York, 1995.

(69) Cativiela, C.; García, J. L.; Elguero, J.; Mathieu, D.; Phan Tan Luu, R. *Quant. Struct.-Act. Relat.* **1987**, *6*, 173–177.

JCTC Journal of Chemical Theory and Computation

Implementation and Optimization of DFT-D/COSab with Respect to Basis Set and Functional: Application to Polar Processes of Furfural Derivatives in Solution

Roberto Peverati and Kim K. Baldridge*

University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received July 15, 2009

Abstract: The implementation, optimization, and performance of DFT-D, including the effects of solvation, has been tested on applications of polar processes in solution, where dispersion and hydrogen bonding is known to be involved. Solvent effects are included using our *ab initio* continuum solvation strategy, COSab, a conductor-like continuum solvation model, modified for ab initio in the quantum chemistry program GAMESS. Structure and properties are investigated across various functionals to evaluate their ability to properly model dispersion and solvation effects. The commonly used S22 set with accurate interaction energies of organic complexes has been used for parametrization studies of dispersion parameters and relevant solvation parameters. Dunning's correlation consistent basis sets, cc-pVnZ (n = D, T), are used in the optimization, together with the Grimme B97-D exchange-correlation functional. Both water ($\varepsilon = 78.4$) and ether ($\varepsilon = 4.33$) environments are considered. Optimized semiempirical dispersion correction parameters and solvent extent radii are proposed for several functionals. We find that special parametrization of the semiempirical dispersion correction when used together in the DFT-D/COSab approach is not necessary. The global performance is quite acceptable in terms of chemical accuracy and suggests that this approach is a reliable as well as economical method for evaluation of solvent effects in systems with dispersive interactions. The resulting theory is applied to a group of push−pull pyrrole systems to illustrate the effects of donor/acceptor and solvation on their conformational and energetic properties.

Introduction

The treatment of van der Waals (vdW, dispersive) interactions is an active field of research as of late, particularly in the DFT community.[1−25] The importance of vdW interactions is clear when considering molecular systems whose structures are largely influenced, and in some cases totally determined, by dispersion interactions. Moreover, the stability of a given molecular charge distribution may also be affected by its environment, specifically by the polarity of the medium or solvent. In particular, aromatic or π-conjugated molecules display a special sensitivity toward the redistribution of electron density as a function of their nuclear position, photonic excitation, and polar solvation. In general, chemical recognition, whether in materials or biological systems, depends on physio-chemical modulation that may be determined either qualitatively or quantitatively. Molecules in close proximity create competition and selection based on relative free energies of reaction (complexation/association), excitation (change in electronic state), and/or product formation (new chemical bond). Proper treatment of dispersion and solvent effects, both important effects on their own, in combination still leave open several important questions for reliable treatment.

Given that dispersion energies are pure electron correlation effects,[26,27] accurate computations via *ab initio* based wave functions, including at least single and double substitutions, can provide good accuracy with adequate basis sets (*e.g.*, CCSD(T) method),[28] albeit with a significant computational cost. On the cheaper side of these methods, Møller−Plesset

* Corresponding author phone: +41 44 635 4201; fax: +41 44 635 6888; e-mail:kimb@oci.uzh.ch.

perturbation (MP2, MP4) methods,[29] while an improvement over most density functional approaches, tend to be largely over bound for many systems.[30,31] In general, neither computational solution is economical, particularly for large molecular systems and complexes. Recent literature including work of our own, on the other hand, illustrates the effectiveness of dispersion corrected DFT approaches.[7,14,15,32–34] Such methods, coupled to effective strategies for treatment of solvent,[35–42] promise to provide an economical yet accurate method where both solvent and dispersion effects dominate. One approach has already been shown in a recent paper by Riley et al.[6] where a slightly modified version of the semiempirical correction of Grimme[14] applied to the TPSS density functional[43] was coupled with the IEF-PCM approach for continuum solvation.[44] However, still not well understood is a more general view of such DFT-D functionals together with continuum solvent representations, including optimal representations of the dispersion correction, effects of basis set, and representations of the solvent.

Our goals in this work include a) contribution to the development of semiempirically corrected density functionals, emphasizing functionals and basis sets that are necessary for investigations of structure and properties in solution, b) implementation, optimization, and performance analysis of the semiempirically augmented density functional theory (DFT-D) in conjunction with our *ab initio* implementation of the conductor-like screening model, COSab, within the GAMESS software,[45] and c) illustration of the combined DFT-D/COSab method on systems with known dispersion and solvation phenomenon, presented with an expanded set of basis sets, functionals, and solvent dielectrics, with consideration of both structure and properties.

Computational Methods

All calculations reported here were carried out using a locally modified version of the GAMESS electronic structure program. In the present work, we consider the semiempirically corrected functional, B97-D,[14] and two conventional hybrid functionals, B3PW91[46,47] and B3LYP.[46,48] Additionally, comparisons are made with the second order Møller–Plesset perturbation theory (MP2)[29] and CCSD(T).[28] Dispersion corrections were recently implemented and tested in the GAMESS software, using the ansatz proposed by Grimme (2006).[16] The B97-D functional is a special reparameterization of the original B97 hybrid functional of Becke.[22] The new parameters in the functional form make B97-D less susceptible to spurious dispersion contamination in the exchange component than the original functional of Becke. The dispersion energy in B97-D is entirely handled by the semiempirical term, leading to smaller errors. For semiempirically corrected functionals, given the importance and dependence of the associated parameters on the choice of functional and basis set, we have carried out parameter optimization for several basis sets, including TZV2P,[49] with (2d,2p), and Dunning's correlation consistent basis sets,[50] denoted cc-pVnZ, where n = D for double with [3s2p1d] contraction, T for triple with [4s3p2d1f] contraction, and Q for quadruple with [5s4p3d2f1g] contraction. The S22

reference set of data provided by Jurecka et al.[51] is used as a validation test set, in addition to several test systems.

Solvation was taken into account using the most recent implementation of our COSab solvation method, based on the original COSMO theory of Klamt and modified for *ab initio* theory.[52–54] Dielectric permittivities of water (ε = 78.4) and ether (ε = 4.335) are used for parametrization studies. An expanded set of dielectrics was used to investigate a set of substituted pyrroles, for illustration of the method, including toluene (ε = 2.38), chloroform (ε = 5.0), t-butanol (ε = 12.0), acetone (ε = 20.7), methanol (ε = 32.6), and water (ε = 78.4). The parameters of the cavity construction are 1082 points for the basic grid and 92 segments on the complete sphere. The outlying charge error was taken into account via the double cavity approach.[53–55] The solvent radial extent was optimized in the parameter optimization studies and taken as 1.3 for the application studies. Solvent atomic radii were taken from Bondi[56] or from Klamt.[57]

Theoretical Approach and Discussion

DFT Dispersion Corrections. Much work has been carried out to revise and enhance density functionals for increased accuracy, the most important component of late being the addition of corrections for dispersion. A particularly effective and consistent strategy to correct failures due to lack of dispersion has involved addition of an empirical potential[19,58] to the final DFT energy. This empirical potential is typically of the form C_6R^{-6}, where R are interatomic distances and C_6 dispersion coefficients. Such a strategy has been formulated and well established by several research groups.[14,15,17,59] In particular, the method of Grimme establishes a semiempirical dispersion correction that can be applied to the final result of any mean field calculation that lacks a sufficient description of dispersion energy.[60,61] Typically, this procedure is used in conjunction with density functional theory (leading to various DFT-D methods)

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

where the dispersion coefficients C_6 are calculated from the atomic polarizabilities, and the f_{dmp} is a damping function with the form

$$f_{\text{dmp}}(R_{ij}) = \frac{1}{1 + \exp\left[-d\left(\frac{R_{ij}}{s_{\text{R}} R_{ij}^0} - 1\right)\right]} \tag{2}$$

This semiempirical correction does not depend on the electronic structure of the system, and, therefore, optimization of the relevant parameters must be considered, in particular the s_6 prefactor to the C_6 atomic coefficients, the d damping factor, and the s_{R} prefactor to the vdW radii. The atomic polarizabilities (which lead to the C_6 coefficients), the vdW radii, and all semiempirical parameters were optimized in the original paper of Grimme,[32] and further work was carried out by our own group more recently.[34] Such parameters have to compensate for differences in the electronic descriptions due to basis sets, functionals, and, in the present work, effects

of solvent. The nonlinear parameter, d, is considered optimal and fixed to the optimized value of Grimme, $d = 20$.[14] Differences in implementations are mainly in the values of these parameters. The final mean-field (MF) DFT-D energy is the result of the direct addition of this dispersion energy to the computed DFT energy

$$E_{MF-D} = E_{MF} + E_{disp} \tag{3}$$

First Principles Continuum Solvation Theory. We have developed a rigorous self-consistent approach for the inclusion of electrostatic solvation effects in conventional *ab initio* gas phase computations in GAMESS.[35,52] This model is implemented at the RHF (ROHF, UHF), DFT, HDFT, and MP2 (UMP2) levels, including first and second derivatives. Our approach to continuum solvent modeling (COSab) arises from the concepts of screening in conductors, a modification of the original COSMO theory of Klamt.[62] Significant model features of this approach include a) molecular-shaped cavity construction, b) efficient solution to energy and derivative quantities, c) efficient ways of including dynamic correlation, d) inclusion of an alternate approach using multipoles up to hexadecapoles, e) assessment of outlying charge effects using a choice of two strategies, f) general electrostatic solvation accuracy within 2−6 kJ/mol, and g) several strategies for inclusion of first solvation shell effects.

We refer the details of the underlying theory of the COSab method to the original paper[52] and instead only point out the main concepts related to the coupling of this theory with the semiempirical dispersion corrections for density functional treatments. The treatment of solvent effects within the continuum model involves the construction of a cavity around a solute system, with a dielectric continuum representing the solvent outside the cavity. The molecular electric field arising from the nuclei and the electronic distribution is screened by the polarization of this dielectric continuum. The iterated solute/solvent interaction can be uniquely represented by the resulting surface charge distribution at the interface to the continuum. Construction of the molecular cavity relies on van der Waals radii of Bondi[56] scaled by a solvent factor (cos_{rad}) in the range of 1.2−1.3 that represents the area of approach of solvent molecules to a solute.[57] These radii in fact ensure that the cavity volumes are approximately equal to the molar volumes of the compounds and therefore are not really free parameters[57,63] and demonstrated the physically correct screening energy decay behavior.[52]

While authors of other solvent models may recommend different sets of vdW radii,[64−66] there is a general agreement that the cavity size is closely related to the vdW surface. Recently, Truhlar et al.[67] provided a systematic study showing variation across four popularly used solvent models, including a discussion of what radii provide the most physical results and also showing that scaling factors greater than ∼1.4 produce unphysical cavities. These important insights in fact imply that, for some molecules (e.g., those with diffuse electronic structure), a small but significant portion of the solute electron density can extend outside the molecular cavity. This leads to error in the prediction of solvent phenomena, since the solute is not fully represented in the proper way in its interaction with the solvent. Such errors

are termed *outlying charge errors* and are most problematic for systems with diffuse charge localized or anionic systems and can lead to errors that are substantial (*e.g.*, 40−80 kJ/mol error). Our present solvation implementation has two effective strategies for the treatment of outlying charge error.[53,68]

It is important to note that the vdW radii involved in the damping function in eq 2 are different than those used in the continuum solvation method, as both are optimized independently using different approaches and for different properties. As such, one cannot use the same set of vdW radii values for both contributions. The model for solvation is in fact independent to that of the semiempirical dispersion model. Any eventual dependencies of the two effects would nevertheless be accounted for through the parametrization of the semiempirical scaling factors within both methods.

Together, the DFT-D+COSab method has a total of 4 parameters to consider for any interdependencies and for optimization. Three parameters, scaling factor s_R and s_6 (vdW radii and dispersion coefficients, respectively), and damping function exponent, d, arise from the DFT-D method. The fourth parameter, cos_{rad}, is related to the creation of the solvent cavity surface. Following the approach used in our previous work (as well as others),[14,34] the value of d is kept fixed at the value of 20. This leaves investigation into the behavior of only three parameters, s_6, s_R, and solvent extent radii (cos_{rad}), simultaneously with respect to basis set and functional.

Parameter Optimization. To perform simultaneous optimization of the three chosen parameters, a suitable database of molecules with accurate solvated interaction energies is required. The S22 set, proposed by Hobza et al.,[51] is the standard database used in all previous considerations of the DFT-D method. However, this database provides accurate interaction energies only in vacuum. To obtain solvated interaction energies, the following term, first proposed by Riley et al.,[6] is used to correct the vacuum energies

$$\Delta E_{solv}^{CP} = \Delta E_{vac}^{CP} + \Delta \Delta E_{solv} \tag{4}$$

In this way, a value for $\Delta \Delta E_{solv}$ can be calculated, using a suitable solvation method, to give

$$\Delta \Delta E_{solv} = \Delta E_{solv}^{no-CP} - \Delta E_{vac}^{no-CP} \tag{5}$$

Using this approach, one can propose a way forward for providing chemical accuracy in solution for a reasonable costs and including important effects of dispersion and solvation. In addition, this method provides a reasonable way to take basis set superposition error (BSSE) into account in solution. This is very important given that continuum solvation methods do not provide a clear and unique way to take into account BSSE in the calculations, since there is no analogue to the counterpoise CP corrections as used in gas phase. Alternatively, since the optimization of the s_6 parameter for the B97-D functional including counterpoise correction shows a minimum at $s_6 = 1.4$ regardless of basis set, one could carry out CP corrected solvation energies using $s_6 = 1.4$.

Optimization of DFT-D including effects of solvent

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2775**

Computational solvent models are fundamentally different at the algorithmic level.[40] This results in different accounting of the various aspects of solvation representation within the quantum mechanics ansatz. For example, models have different (or missing) representations for cavity descriptions, solvent radii, outlying charge effects, nonelectrostatic effects, or other solvent related phenomenon. In particular, comparison with experimental data has shown substantial difference in predictability with different solvent cavity descriptions, which are a function of solvent radii parameters.[69] The radii used to generate the spheres associated with the cavity ultimately govern the volume of the cavity and the distance between the atoms and the solvent surface. Computed properties depend heavily on the choice of atomic radii.

The specific formalism used by Riley et al.[6] involved the use of the Integral-Equation-Formalism Polarized Continuum Model, IEF-PCM, together with the use of united atom, UA0, solvent radii. The general PCM method uses a cavity of interlocking spheres approximately 20% larger than the vdW radius. The electrostatics are treated using the Poisson equation, with Green's functions used to define the integral operators determining the apparent charge. Nonelectrostatic effects are evaluated empirically, using a solvent accessible surface (SAS). In this way, the surface charge depends only on the potential and not on the vector electric field, thereby being less sensitive to numerical instabilities. The specification of the solute cavity, defined using the united atom topological model (UA0), is obtained from spheres centered on non-hydrogen atoms. The radius of each sphere, which is dependent on the atom type, connectivity, and number of hydrogen atoms attached, is based on the Universal Force Field (UFF).[70] This particular choice of radii used for cavity construction (default for cavity construction in other quantum chemistry software) has been found to lead to relatively high mean unsigned errors for solvation phenomenon[69,71] and is typically not recommended for quantitative calculations of solvation free energies.[72–74]

Given the known sensitivities to details of solvation model, it becomes necessary to recalculate solvation reference corrections for the S22 set of molecules for this work. Here, we have calibrated a set of corrections for both high and low dielectric, water ($\varepsilon = 78.4$) and ether ($\varepsilon = 4.33$), respectively. Calculations are performed at the MP2/aug-cc-pVDZ level of theory similar to Riley et al.[6] for comparative purposes. Results are reported in Table 1.

Optimization of the Solvent Extent Parameter (cos_{rad}). The algorithm for cavity construction in the COSab-GAMESS method involves construction of a basic grid representing the molecular cavity generated from the vdW surface around the molecule. This is followed by a series of projections onto the solvent surface, extending the cavity to approximately a distance of van der Waals + 20–30%.[56,57] The cavity segmentation is small enough as to assume homogeneous charge distribution on each segment. The standard value for the parameter, cos_{rad}, is typically taken as 1.2 or 1.3. Here, we actually investigate the response in the range from 1.1 to 1.5 to better understand the influence of this parameter on the prediction of solvation.

Table 1. MP2/aug-cc-pVDZ Solvent Reference Energy Corrections (kJ/mol) for the S22 Set of Molecules

S22 complex reference number	vacuum	$\Delta\Delta E_{solv}$ MP2/COSab (this work)		$\Delta\Delta E_{solv}$ MP2/ IEF-PCM[a]	
		ether $\varepsilon = 4.33$	water $\varepsilon = 78.4$	ether $\varepsilon = 4.33$	water $\varepsilon = 78.4$
1	−13.26	−3.347	2.301	−4.48	−0.54
2	−21.00	−14.81	−11.21	−14.35	−9.87
3	−77.86	−50.58	−36.65	−32.93	−10.42
4	−66.78	−37.53	−20.67	−33.30	−17.40
5	−86.40	−50.71	−30.79	−39.50	−16.44
6	−69.92	−43.30	−28.83	−36.36	−20.88
7	−68.49	−43.76	−31.25	−36.32	−21.67
8	−2.22	−1.97	−1.84	−2.13	−2.13
9	−6.32	−4.35	−3.39	−4.39	−3.68
10	−6.28	−5.10	−4.48	−5.36	−4.64
11	−11.42	−7.11	−5.10	−7.36	−5.23
12	−18.49	−14.39	−11.89	NA[b]	NA[b]
13	−42.34	−23.81	−12.76	−16.40	−7.95
14	−21.84	−14.64	−11.30	−14.85	−12.09
15	−51.17	−34.31	−24.31	−32.09	−24.90
16	−6.402	−2.720	−0.92	−2.55	−0.50
17	−13.72	−7.24	−3.64	0.38	7.03
18	−9.832	−6.99	−5.44	−7.20	−0.08
19	−18.66	−9.83	−4.60	−10.96	−5.56
20	−11.46	−9.54	−8.58	NA[b]	NA[b]
21	−23.97	−16.11	−11.67	NA[b]	NA[b]
22	−29.50	−21.05	−16.78	−15.61	−7.45

[a] Reference 6. [b] Riley et al.[6] did not report this value because of problems in the convergence of the energies.

A first set of calculations was performed at the B97-D+COSab/cc-pVDZ level of theory, with the dispersion parameters fixed at $d = 20$, $s_R = 1.1$, and $s_6 = 1.0$. Further optimization was undertaken of the more critical s_6 parameter for each value of cos_{rad} to confirm results. Figure 1 and Table 2 summarize these results.

The variation in the mean absolute deviation of the S22 set for the cos_{rad} parameter in the range $1.1–1.5$ is within 0.45 kJ/mol for water and 0.13 kJ/mol for ether. An optimal value of $cos_{rad} = 1.30$ is suggested for two reasons. First, the performance is slightly better for the simultaneous optimization of cos_{rad} and s_6 (see black dots on the graphic). Second, 1.30 is one of the possible values suggested for the COSab method (typical values are 1.20 and 1.30). The general message, however, is that values in the range 1.3 ± 0.2 do not affect the general behavior of the method, which could be surprising given the known sensitivity of electrostatic solvation properties to variation in atomic radii, as discussed above.

Optimization of DFT-D Parameters (s_6 and s_R). As in our previous work,[34] we investigate the combined behavior of the two DFT-D parameters, s_6 and s_R, using the optimal value determined for $cos_{rad} = 1.3$. The simultaneous optimization of these two parameters is presented in Figure 2 as a surface function for three different environments, vacuum, water, and ether, was carried out. Results were investigated for the B97-D functional with both the cc-pVDZ and cc-pVTZ basis sets and are illustrated in Figure 2.

The topology of the curves for the three environments (vacuum, ether, H$_2$O) appear very similar to each other but with different relative height on the MAD axis. This similarity in topology suggests that behavior in solvated environments for these complexes with variation in disper-
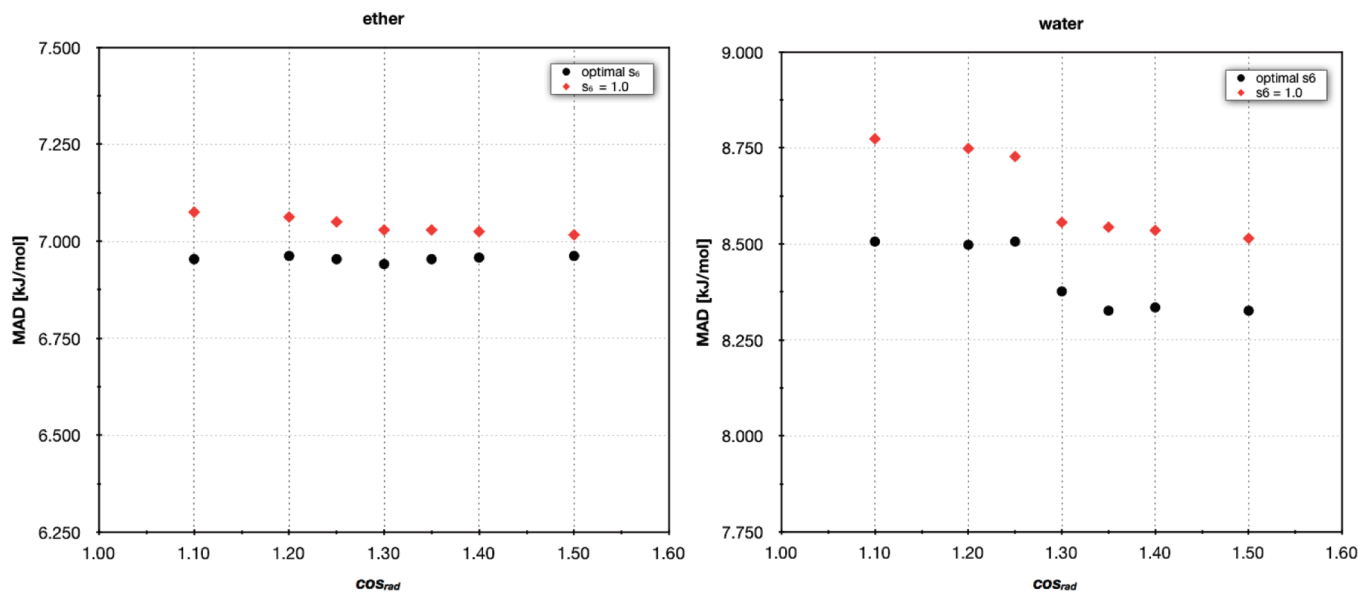
Figure 1. Mean absolute deviation of S22 set of molecules as a function of solvent radii extent value (cos_{rad}) at $s_6 = 1.0$ (red), and at the optimal s_6 for the respective cos_{rad} (black), at the B97-D+COSab/cc-pVDZ level of theory.

Table 2. Summary of Results of Mean Absolute Deviation (MAD, kJ/mol) for the S22 Set Showing Variance in Solvent Radii Extent Parameter, cos_{rad}, as a Function of s_6, at the B97-D+COSab/cc-pVDZ Level of Theory

solvent extent parameter, cos_{rad}	water			ether		
	MAD with $s_6 = 1.0$	corresponding optimal s_6	MAD with optimal s_6	MAD with $s_6 = 1.0$	corresponding optimal s_6	MAD with optimal s_6
1.1	8.774	0.93	8.506	7.075	0.97	6.954
1.2	8.749	0.94	8.498	7.063	0.97	6.962
1.25	8.728	0.95	8.506	7.050	0.97	6.954
1.3	8.556	0.96	8.376	7.029	0.98	6.941
1.35	8.544	0.95	8.326	7.029	0.98	6.954
1.4	8.535	0.95	8.334	7.025	0.98	6.958
1.5	8.514	0.96	8.326	7.017	0.98	6.962
variance	0.01326	-	0.00784	0.00048	-	0.00005
std. error	0.04352	-	0.03346	0.00824	-	0.00270

sion functional parameters is qualitatively the same. The differences regarding the MAD from that of the gas phase are expected due to the fact that these were determined using MP2/aug-cc-pVDZ in the respective environment, rather than at CCSD(T)/CBS. However, because we are interested here in the optimal values of the s_6 and s_R, if the overall topology in different environments is truly within a small tolerance of that observed for the gas phase environment, this difference in height along the MAD axis can be tolerated.

To compare more closely the variance in surface topology in different medium, we can project one curve onto either of the other two and determine any significant point-to-point differences. These differences are plotted as a surface function of the two DFT-D parameters in Figure 3, with a color density map projected on the bottom of each diagram, for all three possibilities: water vs vacuum, ether vs vacuum, and water vs ether. Only the B97-D+COSab/cc-pVDZ case is shown here, since similar results are obtained at the B97-D+COSab/cc-pVTZ level of theory. In these surface representations (Figure 3), blue (negative) means that the solvated curve was lower than the vacuum (ether) curve in Figure 2, while red (positive) means that the solvated curve is higher than the vacuum (ether) curve in Figure 2.

The deviations from zero (red or blue, Figure 3) are in all cases smaller than 4.2 kJ/mol and in general very small for both basis sets, confirming the fact that the curves overlap almost perfectly. Moreover, the intersection of the optimized s_6/s_R parameters is always in a minimum deviation region. From these results, we can conclude that the method behaves in the same way for both solvents, when the optimization of the two DFT-D parameters is conducted simultaneously. When solvated results are compared to the vacuum case, this general conclusion does not change.

A more detailed slice of the surface functions is shown in Figure 4 for $s_R = 1.1$. Numerical values for the corresponding s_6 optimizations are reported in Table 3. As can be seen, the values of the optimal s_6 parameter for the solvated cases differ only slightly from that used in the vacuum, for each of the basis sets considered. Additionally, these results also include indirectly, consideration of BSSE, as shown above. Therefore, our overall conclusions made in our previous article, including discussion of BSSE, more extensive optimization, and summary of values proposed as a function of basis set, can be extended with good accuracy to the DFT-D + COSab procedure, for calculation of interaction energies in solution. As such, these results, together with
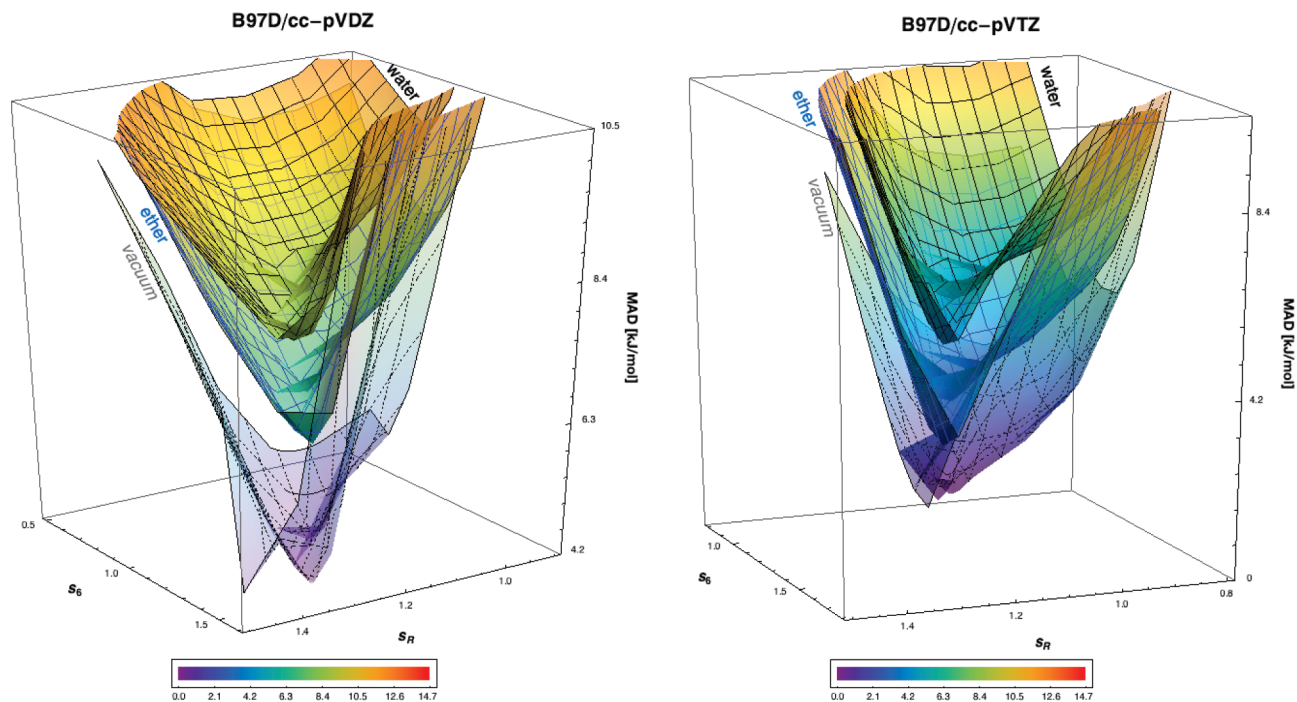
Optimization of DFT-D including effects of solvent

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2777**

Figure 2. Optimized s_6 and s_R empirical dispersion function parameters for the B97-D functional, with cc-pVDZ and cc-pVTZ basis sets, respectively. Plots are given on the same relative scale.
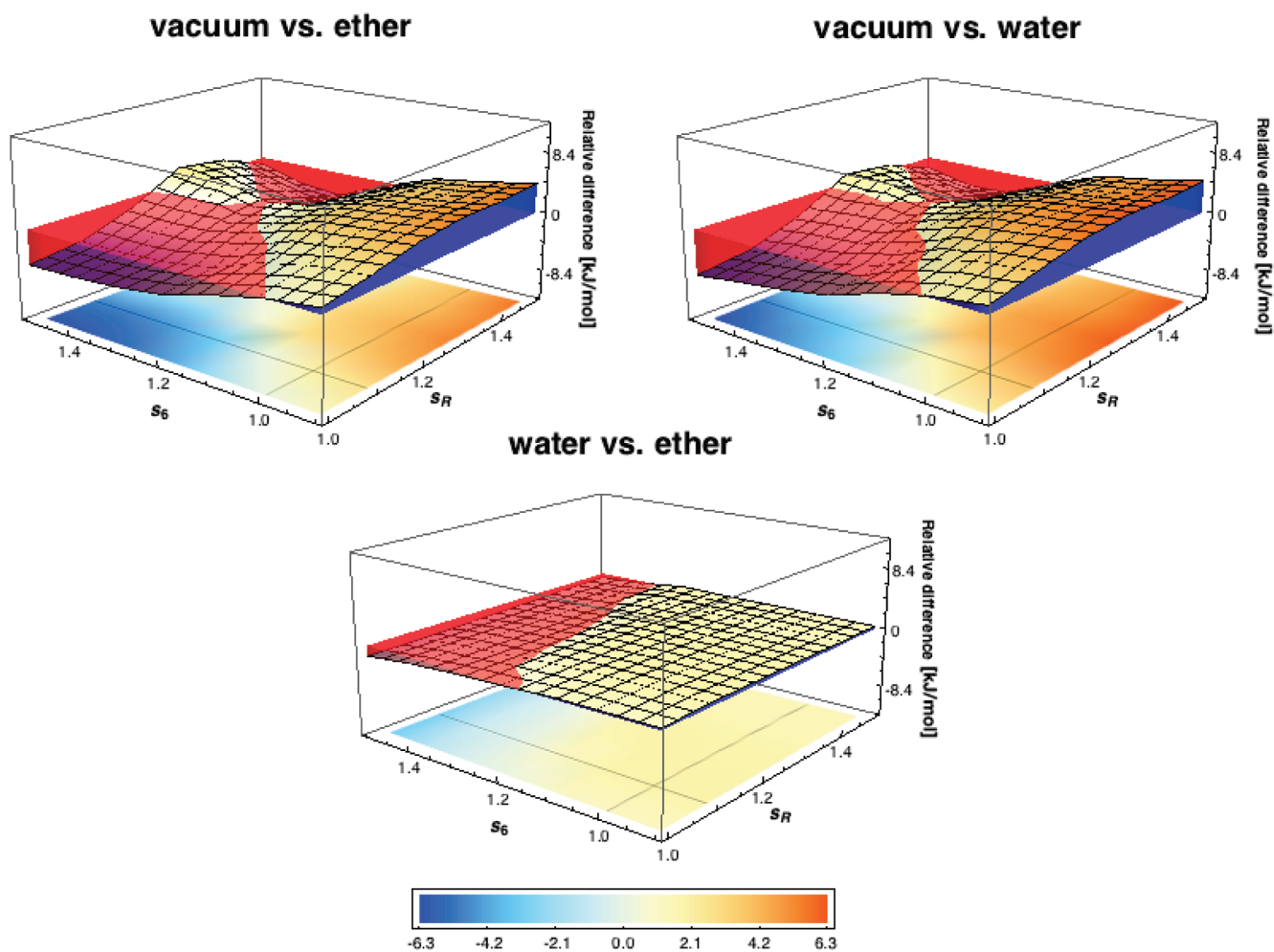

Figure 3. Water vs vacuum, ether vs vacuum, and water vs ether projection plots, at the B97-D+COSab/cc-pVDZ level of theory.
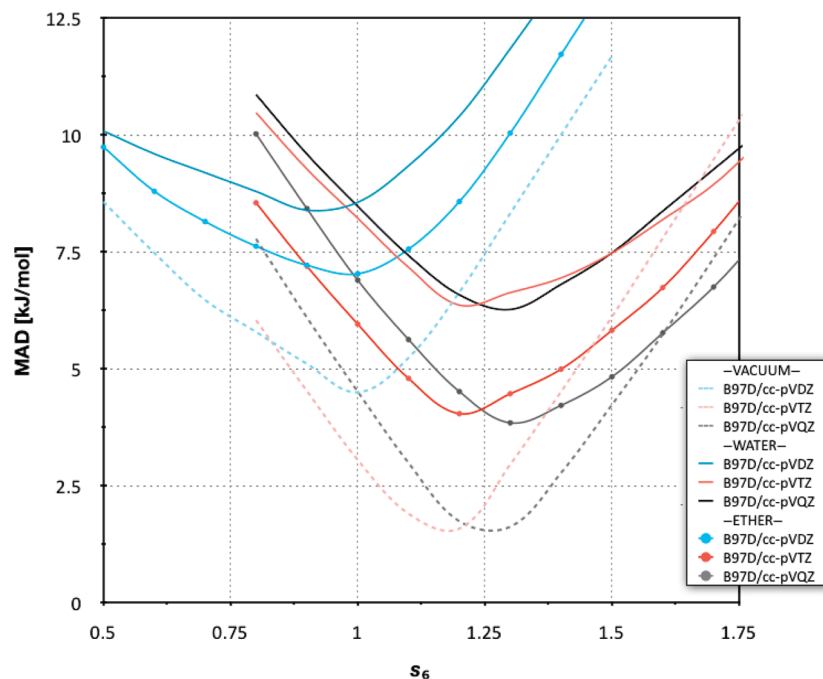
Figure 4. Water vs vacuum, ether vs vacuum, and water vs ether projection plots at the B97-D+COSab/cc-pVDZ level of theory, represented as a slice through the surface functions from Figure 3.

Table 3. Summary of Results of Mean Absolute Deviation (MAD, kJ/mol) for the S22 Set Showing Variance in s_6 with Solvent and Basis Set[a]

solvent	basis-set	optimized s_6	MAD [kJ/mol]
vacuum	cc-pVDZ	1.00	4.498
	cc-pVTZ	1.18	1.410
	cc-pVQZ	1.26	1.381
water	cc-pVDZ	0.96	8.376
	cc-pVTZ	1.21	6.356
	cc-pVQZ	1.28	6.192
ether	cc-pVDZ	0.98	6.941
	cc-pVTZ	1.19	4.038
	cc-pVQZ	1.28	3.820

[a] $cos_{rad} = 1.3$, $s_R = 1.1$, $d = 20$.

those of Riley et al.[6] using a different ansatz for including solvation effects together with semiempirically corrected DFT, support that in general, for any treatment of dispersion-governed systems in solution, no additional considerations need be made nor specialized optimizations undertaken.

Geometry Optimization. The DFT-D method is typically parametrized using single point energy calculations on an accurate gas phase geometry, for each complex in the S22 database. This procedure opens up questions regarding any variations in methodology and subsequent prediction of results, when full optimization of geometry is considered instead of a single point energy calculation. In the case of a vacuum environment, these questions are not considered significant due to the fact that the starting geometry is calculated at a very high level of theory. More specifically, the optimized geometry for most of the complexes in the S22 set has been determined with CCSD(T)/QZ, with a few cases optimized with CCSD(T)/TZ or MP2/TZ-CP. All final single point energies are then calculated at the CCSD(T)/CBS level. Explicit details of the optimizations used in each case can be found in the original paper.[51]

In contrast, it is not readily clear that the gas phase optimized geometries are the most appropriate when one is considering a solvent environment, as one can expect changes in the structures of the complexes due to their interactions with the solvent. In particular, one expects a change in distance between the two monomers in the complex. While the main purpose of this article is not to search for better procedures of parametrization of the DFT-D method, one still must consider possible sources of error due to changes in geometry and the consequences for the DFT-D + COSab procedure.

In order to carry out such an analysis, we have calculated the gradient for all molecular systems in the S22 set with the B97-D functional across a variety of basis sets, in a water environment. The root-mean-square deviation (RMSD) and the largest component (MAX) of the gradient are reported in Table 4, as a measure of the differences between the geometries as given in the S22 set and their respective optimized equilibrium geometries.

These summarized values for rmsd and MAX gradient indicate that, provided one uses a large enough basis set, a large majority of the complexes are not far from their gas phase optimized geometries. As expected, an incremental increase in basis set size provides a more accurate picture of just how close the solution phase complexes are to the optimized gas phase geometry. Cases where there are the greatest rmsd values, an indication of when solvation changes are expected to be significant, are highlighted in boldface in Table 4. For example, complexes that have bond length differences in the picometer range represent significant differences when one is trying to obtain chemical accuracy in structure and associated properties.

To further analyze these results, complexes in the S22 set were divided into three subgroups according to the nature of the primary interactions between the molecules. The three

Table 4. Summary of Results of RMSD and Maximum (MAX) Gradient Values for the S22 Set As a Function of Basis Set

molecule	cc-pVDZ		cc-pVTZ		cc-pVQZ	
	RMSD gradient	MAX gradient	RMSD gradient	MAX gradient	RMSD gradient	RMAX gradient
1	0.0077984	0.0174154	0.0026740	0.0053823	0.0022662	0.0041095
2	**0.0105507**	0.0191442	0.0056189	0.0124492	0.0051123	0.0120283
3	0.0083584	0.017122	0.0046930	0.0097907	0.0040162	0.0090413
4	0.0090051	0.0234469	**0.0062381**	0.0166286	**0.0058922**	0.0159450
5	0.0078098	**0.0249773**	0.0055591	**0.0178010**	0.0053169	**0.0165761**
6	0.0058952	0.0203691	0.0036121	0.0149223	0.0035065	0.0139418
7	0.0068203	0.0221163	0.0044578	0.0146959	0.0042735	0.0130864
8	0.0055312	0.0107665	0.0020459	0.0040859	0.0018506	0.0037130
9	0.0054494	0.0098256	0.0025617	0.0035783	0.0024370	0.0034128
10	0.0052723	0.0124369	0.0022704	0.0058819	0.0021843	0.0056609
11	0.0044198	0.0094194	0.0016463	0.0035123	0.0015786	0.0032221
12	0.0047943	0.0101749	0.0019266	0.0039867	0.0021137	0.0040012
13	**0.0075948**	**0.0222007**	**0.0055311**	**0.0150658**	**0.0052161**	**0.0132548**
14	0.0042266	0.0099944	0.0017430	0.0037012	0.0016530	0.0037704
15	0.0061554	0.0210383	0.0038124	0.0139875	0.0036075	0.0122721
16	**0.0066904**	0.0164377	0.0025598	0.0047258	0.0026196	0.0057154
17	0.0054453	0.0134867	0.0023467	0.0076936	0.0021896	0.0070529
18	0.0057686	**0.0179676**	0.0022156	0.0057129	0.0020522	0.0045684
19	0.0049909	0.0141728	**0.0063532**	**0.0313533**	**0.0068787**	**0.0340759**
20	0.0044762	0.0109458	0.0017516	0.0041108	0.0016882	0.0038493
21	0.0043044	0.0105029	0.0018935	0.0044323	0.0018214	0.0041939
22	0.0049381	0.0122261	0.0022128	0.0071765	0.0021013	0.0065132
average	0.0061953	0.0157358	0.0033511	0.0095761	0.0031989	0.0090911
maximum value	0.0105507	0.0249773	0.0063532	0.0313533	0.0068787	0.0340759
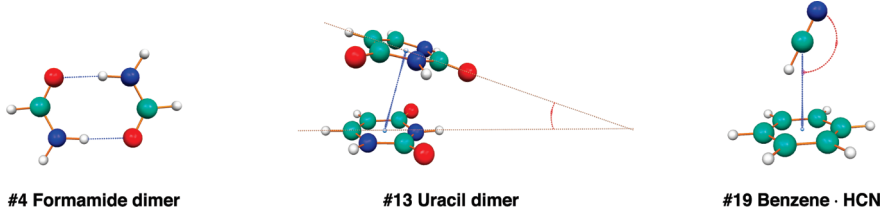
categories of complexes are A) hydrogen-bonded complexes (1−7), B) predominant dispersion interactions, complexes (8−15), and C) mixed hydrogen bonding and dispersion, complexes (16−22). Complexes with the highest rmsd and MAX gradient within each group are indicated in boldface in Table 4. In each category, we have chosen one complex to illustrate results of a full optimization in solution at the B97-D/cc-pVnZ (n = D,T,Q) level of theory, starting from the accurate gas-phase geometry. In category A, H-bond complexes, formamide dimer (#4) is chosen, with the highest rmsd gradient at cc-pVTZ and cc-pVQZ, and the second highest at cc-pVDZ. For category B, the dispersion complexes, the uracil dimer stack (#13) is the complex with highest rmsd and MAX gradient with all considered basis sets. Within category C, the mixed complexes, benzene-HCN (#19) has the highest rmsd and MAX gradient with cc-pVTZ and cc-pVQZ and also for cc-pVDZ. Data for the fully optimized solvent (water) structures and energetics for these three complexes as compared to gas phase are illustrated and summarized in Table 5.

The most sensitive geometry parameters in these complexes are the intermonomer distances and the angular relationship between the two monomers. Table 5 summarizes these values as a function of basis set and environment for the three complexes considered. The geometry taken from the S22 set has been optimized in the gas phase and in water for the basis sets shown. In the gas phase, each complex shows a different behavior in the variance of the interatomic distance with respect to increase in basis set. The intermonomer distance in the formamide dimer increases significantly with increase in basis set, with a convergence to the S22 reference value (ΔReference = 0.032 Å, 0.001 Å, and 0.005 Å, respectively). Complex 13, uracil dimer, on the other hand, shows a decrease in intermonomer distance with increase

in basis set, contrary to the reference limit (ΔReference = 0.008 Å, 0.021 Å, and 0.086 Å, respectively). Addition of diffuse functions does not change this trend. The angular relationship between the two monomers is, however, similar to that predicted at the reference level. Complex 19 shows an increase in intermonomer distance with basis set but away from the reference value of 3.387 Å (ΔReference = 0.031 Å, 0.024 Å, and 0.043 Å, respectively). The angular dependence in the benzene-HCN system is relatively the same across basis sets, with a total variance of ∼5°.

Moving from gas phase to solution phase results in a significant change in geometry but again different trends for the three different cases. In the case of the formamide dimer, the hydrogen bond distance increases by almost 6 pm in going from the gas to the solution phase. It is known that high dielectric solvents have a dramatic effect in the dimerization of formamide.[75−77] The increase in the polarity of the solvent is thought to reduce the dimerization, likely due to competition from the surrounding medium. This then causes the opportunity for other bonded forms, from the rather well-defined double hydrogen bonded form in the gas phase, to single hydrogen-bond dimers (medium dielectric) and less well-defined structures in water.

In the second class of complexes, the uracil dimer, the change in environment from gas phase to solution phase shows an even more dramatic increase in the distance between the two monomers (*e.g.*, ∼15 pm for cc-pVQZ). The uracil dimer can exist in both hydrogen-bonded configurations as well as various stacked arrangements.[78] The stacked and T-shaped conformations, with a large dispersive energetic component, are less stable than the hydrogen bonded configurations, albeit less so in the gas phase than in the solution phase.

Table 5. Comparison of Optimized Vacuum, Optimized Solvent (Water), and Reference S22 Gas Phase, Geometry Parameters for Three Complexes of the S22 Set Using Different Basis Sets[a]

| | #4 Formamide dimer | #13 Uracil dimer | #19 Benzene · HCN |

bonds	S22 reference CCSD(T)/CBS	vacuum			water		
		cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pVDZ	cc-pVTZ	cc-pVQZ
#4	**1.841**	1.809	1.842	1.846	1.860	1.901	1.905
#13	**3.365**	3.357	3.344	3.279	3.516	3.459	3.434
#19	**3.387**	3.418	3.412	3.430	3.279	3.226	3.180

angles	S22 reference CCSD(T)/CBS	vacuum			water		
		cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pVDZ	cc-pVTZ	cc-pVQZ
#4	-	-	-	-	-	-	-
#13	**13.8**	18.4	18.0	14.0	19.9	13.6	12.5
#19	**174.6**	175.7	177.0	179.6	154.3	153.7	149.4

	respective optimized vacuum geometry vs optimized solvated (water) geometry								
	initial solvated rms (au) from gas phase geometry			ΔE (kJ/mol)$_{vacuum-water}$			%max geometry change (vacuum-water)		
	cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pVDZ	cc-pVTZ	cc-pVQZ
#4	0.00905	0.00624	0.00589	1.46	4.595	3.93	1.1	3.8	3.2
#13	0.00760	0.00553	0.00522	10.70	4.51	2.50	5.9	1.6	0.3
#19	0.00499	0.00635	0.00688	16.74	19.30	18.28	8.0	16.5	3.4

[a] The graphic shows intermolecular distances (blue) and angles (red), as reported in the table.

In the last case, benzene-HCN system, one sees a very large decrease in the intermonomer distance in the high dielectric water environment, changing from 3.43 Å to 3.18 Å, at the B97-D/cc-pVQZ level of theory. In addition, the angular relationship between the benzene and the HCN changes significantly, from perpendicular in the gas phase, to canted in the solution environment. Other studies have also shown that this interaction weakens with stronger dielectric, with the interaction energy in water being less than half as strong as in the gas phase (~15 kJ/mol vs ~6.5 kJ/mol).[6]

In all of these cases, taking the optimized gas phase geometry as the starting point of the solution phase optimization, one sees that the gradient of the first step is quite high (on average 0.006 au), an indication that the solution phase structure is still far from the gas phase structure. For the three molecules considered here, one sees maximum geometry changes as high as 16% from that of the gas phase, depending on the system and level of theory. Such differences between gas and solution phase can constitute significant differences in energetics, for example, if one is trying to determine interaction energies in solution environment with a high level of accuracy but is only using gas phase geometries. In just these three examples, we see over 17 kJ/mol in energy difference (at the cc-pVQZ level of theory) between gas and solution phase. In general, one must carefully consider the implications of geometry optimization

in a solution environment depending on the system and associated properties of interest.

Illustrative Example: Ground State Conformational Dynamics of Polar Processes

Despite the general tendency for molecules to adopt electronic configurations with minimal charge separation (e.g., electroneutrality principle), many important structural and dynamical features are mediated by charge-separated states. In particular, the conformational analysis of push−pull conjugated π systems is strongly influenced by contributions from zwitterionic states. In donor−acceptor systems, one can identify a dual-resonance-form model, where the major contributor is determined by the degree of stabilization of the charge-separated state. Typically, one of the two resonance forms contributes local double-bond character, manifesting a higher rotational barrier. Depending on the nature of the donor (D)/acceptor (A) pair and the effect of solvent, the barrier to rotation can span a fairly large range. In addition to kinetic effects, it is often found that the dominant equilibrium conformer varies with the medium, with the conformer of higher dipole moment generally more favored in media of high dielectric constant.

A particular case of longstanding interest is the barrier to rotation in push−pull furfural systems, especially due to their
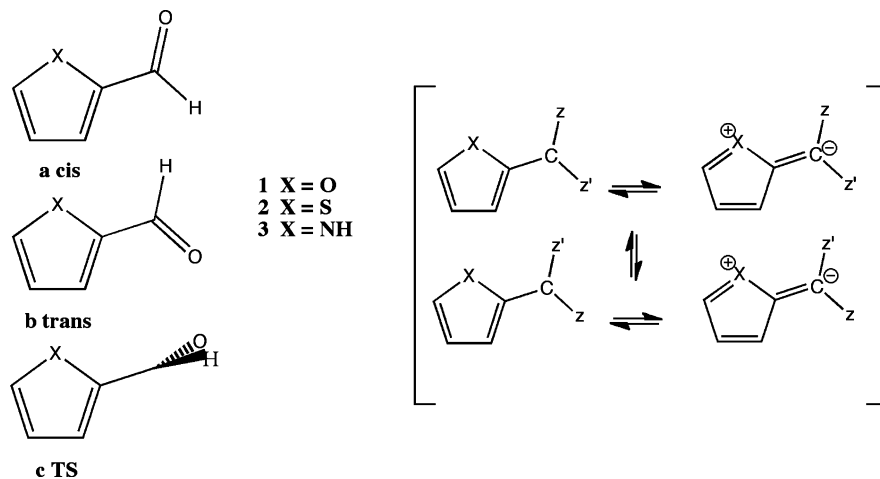
Figure 5. Conformations of furan-2-carbaldehyde (X = O), thiophen-2-carbaldehyde (X = S), and pyrrole-2-carbaldehyde (X = NH).
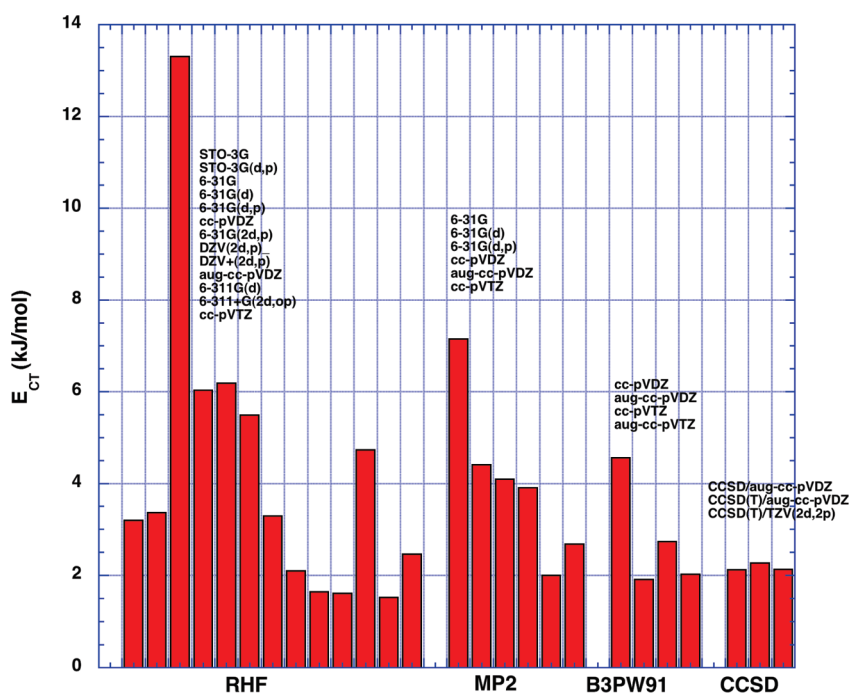

Figure 6. Rotational equilibrium of *trans-* vs *cis-*2-furanaldehyde as a function of wave function type and basis set, relative to the *trans* isomer.

importance as components in many biological redox centers and synthetic molecular devices.[79−84] Moreover, these systems have associated experimental as well as computational challenges.[85] For example, the very small energy difference between the *cis* and *trans* isomers, the possibility for secondary interactions in some of the analogues as well as the strong solvent dependencies across a small range of dielectric (particularly for the parent furan-2-carbaldehyde system), creates particular challenges for both experimental as well as computational predictions. Unambiguous assignments of experimental spectral results have made difficult the predictions of the *cis-trans* energy difference, ΔE_{CT}, as well as the barrier to interconversion, E_a, with a wide range in experimental predictions in both (see, e.g., the Supporting Information).[86] Similar details pose challenges for computational solvent methods to accurately predict the small energy differences between isomers and pick up the solvent

detail that reproduces the conformational change as a function of dielectric. Solvent effects can actually be large enough to reverse the order of stability between the *cis* isomer and the *trans* isomer. As such, this class of system represents a particularly good test for the present method. In particular, here we investigate the parent furfural system and two analogues, as illustrated in Figure 5.

In our 2000 work, together with the experimental group of Bain,[87] we investigated a full set of push−pull conjugated pyrrole cognates. Bain and Hazendonk[87] used a combination of three NMR experiments to obtain rate data over 6 orders of magnitude representing an approximately 150 K temperature range, giving very reliable experimental numbers for the furfural system. Experimental activation parameters were obtained with errors less than 1 kJ/mol and 6 J/(mol K) for $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$, respectively. The $\Delta H^{\ddagger}$ for toluene, acetone, and methanol do not obey a simple relationship with ε,

Table 6. Rotational Equilibrium (kJ/mol) of *trans-* vs *cis*-2-Furanaldehyde As a Function of Wave Function Type and Basis Set, for Gas Phase and 6 Different Solvents[a,d]

	cc-pVDZ				aug-cc-pVDZ			
ε	MP2	B97-D	B3LYP	B3PW91	MP2	B97-D	B3LYP	tB3PW91
1[b]	3.72	4.56 (4.61)	4.82 (4.93)	4.65 (4.75)	2.02	1.92 (2.31)	2.03 (2.50)	1.94 (2.40)
2.38	2.29	3.19 (3.35)	3.15 (3.37)	2.98 (3.13)	0.03	−0.10 (1.28)	0.49 (1.07)	−0.13 (0.38)
5	1.29	1.96 (2.09)	1.60 (1.86)	2.17 (2.37)	−1.47	−1.56 (−1.16)	−1.78 (−1.17)	−1.73 (−1.13)
12	0.55	1.08 (1.16)	1.17 (1.43)	1.01 (1.21)	−2.57	−2.72 (−1.78)	−2.22 (−1.57)	−3.47 (−2.87)
20.7	0.29	1.00 (1.10)	0.64 (0.83)	1.65 (1.92)	−3.06	−3.06 (−1.77)	−3.41 (−2.76)	−3.87 (−2.03)
32.6	0.15	0.88 (0.98)	0.66 (0.86)	0.65 (0.86)	−3.23	−3.44 (−2.12)	−3.83 (−3.15)	−3.94 (−3.36)
78.4	0.02	0.56 (0.65)	0.88 (1.09)	0.46 (0.69)	−3.45	−3.74 (−2.19)	−3.09 (−2.39)	−3.98 (−3.35)
average unsigned difference with respect to MP2	-	−0.70 (−0.80)	−0.66 (−0.87)	−0.75 (−0.94)	-	−0.14 (−0.90)	−0.35 (−0.61)	−0.36 (−0.43)
difference with respect to CCSD(T)[c] for $\varepsilon = 1$	1.59	2.43 (2.48)	2.69 (2.80)	2.52 (2.62)	−0.11	−0.21 (0.18)	−0.10 (0.37)	−0.19 (0.27)
difference with respect to experiment for $\varepsilon = 1$	0.30	1.14 (1.19)	1.40 (1.51)	1.23 (1.33)	−1.40	−1.50 (−1.11)	−1.39 (−0.92)	−1.48 (−1.02)

[a] ΔE_{CT} relative to the *trans* isomer **1a** in kJ/mol. [b] Experimental gas phase value reported as 3.42 ± 0.38.[88] [c] CCSD(T)/DZV(2d,p)//MP2/DZV(2d,p) [$\Delta E_{CT} = 2.99$, $E_a = 42.91$ kJ/mol], CCSD(T)/TZV(2d,2p)//MP2/TZV(2d,2p) [$\Delta E_{CT} = 2.13$, $E_a = 39.16$ kJ/mol], and CCSD(T)/aug-cc-pVDZ//B3PW91/aug-cc-pVDZ [$\Delta E_{CT} = 2.26$, $E_a = 41.10$ kJ/mol].[86] [d] Zero point corrected values are given in parentheses.

Table 7. B97-D/aug-cc-pVDZ Dipole Moments of *trans*- and *cis*-Conformers for 2-Furanaldehyde, **1**, in Gas Phase and 6 Different Solvents[b]

	2-furanaldehyde			
	dipole (D)			
ε	cis	trans	TS	ΔE_{CT},[a] kJ/mol
1	4.46	3.85	2.82	−1.92 (−2.31)
2.38	5.29	4.57	3.24	0.10 (−1.28)
5	5.87	5.07	3.52	1.56 (1.16)
12	6.31	5.45	3.72	2.72 (1.78)
20.7	6.47	5.59	3.79	3.06 (1.77)
32.6	6.55	5.66	3.82	3.44 (2.12)
78.4	6.64	5.74	3.86	3.74 (2.19)
Δ(gas−water)	2.18	1.89	1.04	5.66 (4.50)

[a] ΔE_{CT} relative to the *trans* isomer **1a** in kJ/mol. [b] Zero point corrected values are given in parentheses.

Figure 7. Comparison of calculated ΔE_{CT} as a function of dielectric for various wave function types using the aug-cc-pVDZ basis set. The insert figure shows an enhanced view with additional calculated points of the *x* axis region from $\ln(\varepsilon) = 2.5-3.5$.

suggesting that the solvent effect is likely more complex than just the effect of solvent reaction field. Large $\Delta S^{\ddagger}$ values support this and also implied that equating $\Delta G^{\ddagger}$ and $\Delta H^{\ddagger}$, as is often done, is not justified even for aprotic solvents. Using theoretical methods, we previously carried out a full investigation[86] of the effects of method/basis set on prediction of both ΔE_{CT} and barrier to interconversion. A summary of these results is shown in Figure 6. Results enable comparison across basis set as well as wave function type. Basis sets of at least double-ζ quality with polarization on both heavy and light atoms, and preferably diffuse functions, is shown to be optimal. Our best estimate for ΔE_{CT} (relative to the lower energy *trans* form) and barrier was determined with CCSD(T)/TZV(2d,2p)//MP2/TZV(2d,2p) [$\Delta E_{CT} = 2.13$, $\Delta E_{CT\text{-}ZPE} = 2.48$, $E_{a\text{-}ZPE} = 37.62$ kJ/mol]. This provided benchmark values, given the historical controversy regarding the experimental information. However, less expensive but with overall good predictability is the MP2/aug-cc-pVDZ level. Theoretical predictions can be compared to the 1989 IR and Raman spectroscopic study reporting $\Delta E_{CT} = 3.42 \pm 0.28$ kJ/mol and $E_a = 38.96 \pm 0.25$ kJ/mol.[88]

We now compare the earlier, more computationally expensive CCSD(T) and MP2 results, with the present dispersion enabled functional, B97-D, for analogues **1**−**3** in

both gas and solution phase. We first consider the parent, furan-2-carbaldehyde system (furfural), **1**. The energy difference between *trans*, **1a**, and *cis*, **1b**, forms of furfural in gas phase shows the *trans* conformation, **1a**, to be more stable by ∼4 kJ/mol, with the exact value dependent on the level of theory, as previously found. In solution, the larger dipole moment of the *cis* isomer, **1b**, can preferentially stabilize this form over the *trans* conformation. In Table 6, results with B97-D/cc-pVDZ and B97-D/aug-cc-pVDZ are compared to other functionals as well as our previous MP2 and reference CCSD(T) results, across a wide range of dielectrics. Gas phase ΔE_{CT} values are overestimated with cc-pVDZ with all methods with respect to the benchmark CCSD(T) values. However, the aug-cc-pVDZ basis provides sufficiently accurate results for all considered functionals. The average unsigned difference with respect to MP2 shows that B97-D provides better results than B3LYP or B3PW91, at an overall lower computational cost than for hybrid functionals, MP2, or CCSD(T).

Table 8. B97-D/aug-cc-pVDZ Calculated Gas Phase Energetics and Dipole Moments for **1**−**3**

		energetics[a]						dipole		
X	ΔE_{CT}	expt	preferred tautomer in gas phase	E_a	expt	cis	trans	TS	preferred tautomer in solution phase	expt
O	1.9	3.42 ± 0.28[b]	*trans*	55.2	38.96 ± 0.25[b]	4.5	3.8	2.8	*cis*	3.97 (μ_{cis})[c]
										3.52 (μ_{trans})[c]
										3.52 (μ_{TS})[c]
S	−3.85	4.1 ± 0.4	*cis*	45.5	-	4.1	3.7	2.8	*cis*	3.52 (μ_{cis})[d]
NH	−14.7	-	*cis*	57.1	-	3.1	4.6	2.7	*trans*	2.48 (μ_{cis})[e]

[a] ΔE_{CT} relative to the *trans* isomer, **1a**, in kJ/mol. [b] Reference 88. [c] *cis* and *trans* from ref 89. TS value calculated from bond dipoles, ref 90. [d] Reference 91. [e] Reference 92.

Table 9. B97-D/aug-cc-pVDZ Dipole Moments of *trans*- and *cis*-Conformers for Thiophen-2-carbaldehyde, **2** (X = S), and Pyrrole-2-carbaldehyde, **3** (X = NH), in Gas Phase and 6 Different Solvents[a]

	S					NH				
ε	cis	trans	TS	preferrence wrt dipole in solution	ΔE_{CT}	cis	trans	TS	preferrence wrt dipole in solution	ΔE_{CT}
1	4.08	3.68	2.82	*cis*	3.85 (4.41)	3.07	4.64	2.74	*trans*	14.7 (14.0)
2.38	4.88	4.39	3.26	*cis*	4.28 (4.96)	3.72	5.54	3.19	*trans*	11.7 (11.5))
5	5.44	4.89	3.56	*cis*	4.66 (5.35)	4.18	6.19	3.49	*trans*	9.36 (9.17)
12	5.87	5.27	3.78	*cis*	5.10 (5.82)	4.53	6.67	3.72	*trans*	7.28 (7.21)
20.7	6.02	5.41	3.86	*cis*	5.06 (5.90)	4.66	6.84	3.79	*trans*	5.37 (5.40)
32.6	6.10	5.48	3.90	*cis*	5.19 (5.97)	4.72	6.94	3.83	*trans*	6.26
78.4	6.20	5.57	3.95	*cis*	4.84 (5.73)	4.80	7.04	3.87	*trans*	4.48 (4.33)

[a] ΔE_{CT} relative to the *cis* isomer, **1b**, in kJ/mol.

The general trend across the range of dielectrics (natural log scale), from toluene (ε = 2.38) to water (ε = 78.4) for the parent system, for a variety of wave function types, is shown in Figure 7. Notably, the B97-D and MP2 curves are very similar in predictability, while the two hybrid DFT methods show considerable variance, particularly the B3LYP method. We have carried out additional computations for the B3LYP curve around the region of ε = 5.0, and, in fact, the prediction is shown to be continuous around these points, rather than exceptions (see, *e.g.*, Figure 7 insert graph). This region corresponds to the cross over between preferences for *cis* over *trans* conformation. The overall dipole moment change over the span of dielectrics is $\Delta\mu$ = 1.9 D for a corresponding change in $\Delta\Delta E_{CT}$ = 4.50 kJ/mol. The slope of the curve on the ln(ε) plot is relatively steep illustrating the strong change preference toward the *cis*-2-furanaldehyde form over the *trans*-2-furanaldehyde form in higher dielectric. While it is true that the *trans*-2-furanaldehyde form relieves the lone-pair/lone-pair oxygen repulsion that is found in the *cis*-2-furanaldehyde form, in higher dielectric the higher dipole moment of the *cis* form relative to the *trans* form will eventually dominate, changing that preference.

To further illustrate donor/acceptor properties in these systems, we have investigated thiophen-2-carbaldehyde (X = S) and pyrrole-2-carbaldehyde (X = NH). Table 8 summarizes the gas phase results for the three analogues. Unlike the parent system, thiophene-2-carbaldehyde, **2**, has a more polar S-*cis* conformation and is therefore more stable in the gas phase than is the S-*trans* conformation. Although the dipole difference between S-*cis* and S-*trans* is smaller than in **1**, one still expects preferential stabilization of the former conformation in solvent. Similarly, pyrrole-2-carbaldehyde, **3**, is also found to have a more stable *cis*-conformation in the gas phase. Importantly, the less polar N(H)-*cis* form gains stabilization due to the possibility for

intramolecular interaction between the hydrogen of the nitrogen and the oxygen of the furan ring. In this case, however, unlike either **1** or **2**, the dipole moment would suggest a substantial preferential stabilization of the *trans* conformation in solution environment, due to the significantly larger dipole of the NH-*trans* structure, so the two effects, weak interactions in the *cis* form and strong dipole stabilization in the *trans* form, are competitive.

The tautomeric equilibrium in solution phase is in general found to be dependent on both ΔE_{CT} and $\Delta\mu$. Table 9 summarizes the effects of dielectric variance for analogues, **2**, S, and **3**, N(H). One sees a consistent preference for the *cis* isomer in both the calculated ΔE_{CT} as well as the predicted dipole across the full span of dielectrics. In **3**, N(H), as observed in the gas phase results, we see the competition between the hydrogen bond stabilization present in the *cis* isomer, and the stronger dipole component present in the *trans* isomer. The result is a preference for the *cis* isomer across the span of dielectric, since the weak hydrogen bond interaction will only partially compensate for the more dramatic solvent dipole effect in the N(H)-*trans* conformation.

A subtle feature of the solvation model capability is observed with system **2**, S, due to the second row element, sulfur. In particular, the detailed nature of the cavity construction algorithm is an inherent part of the quality of the solvation model. Different regions of the molecular cavity can have different sensitivity to surface discretization, depending on how rapidly varying the surface charge is, and may require a higher level of discretization to track this change properly. In the case of thiophen-2-carbaldehyde, one can observe this phenomenon. Figure 8 shows a comparison of ΔE_{CT} as a function of ln(ε) for a normal cavity discretization and a highly refined cavity discretization. Checking also the effects of outlying charge and zero point energy, one sees that each of these have a consistent
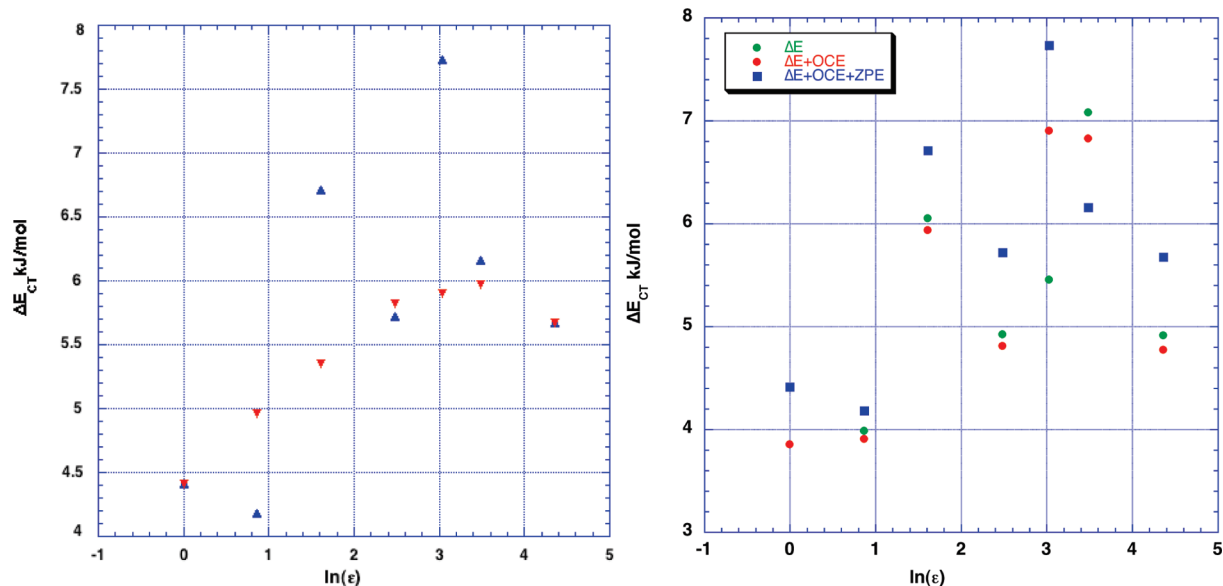
Figure 8. Variation in ΔE_{CT} as a function of dielectric for two levels of cavity discretization (a) and as a function of outlying charge and zero point energy (b) for thiophen-2-carbaldehyde, **2**.
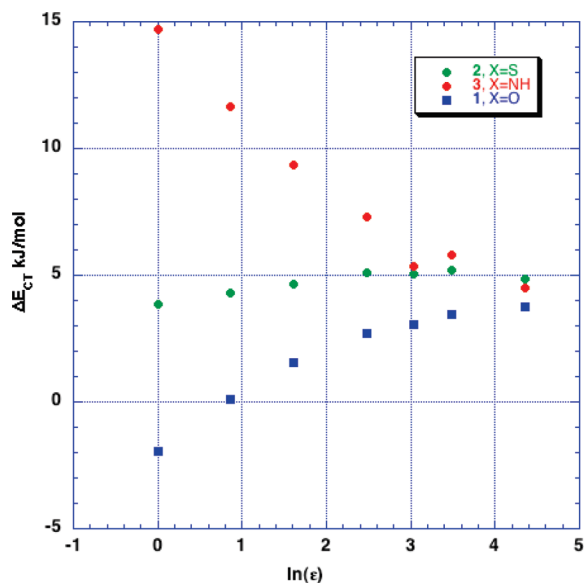

Figure 9. Comparison of calculated ΔE_{CT} as a function of cavity discretization (a) and as a function of outlying charge and zero point energy (b) for thiophen-2-carbaldehyde, **2**.

Figure 10. Comparison of calculated E_a (*cis*→TS) as a function of dielectric (ln(ε)).

effect across dielectric and are not responsible for the inconsistent trend in the data.

Figure 9 shows comparison of solvation trends for all three analogues at the B97-D/aug-cc-pVDZ level of theory. The curves illustrate the dipole versus electrostatic trends across the span of dielectrics for the different types of functionalities. In particular, pyrrole-2-carbaldehyde **3**, shows a steep slope in the opposite direction as that of the parent furfural system, for the reasons explained above. In the sulfur substitution, the smaller change in dipole over the span of dielectrics, results in only a small increase in stabilization energy across the span of dielectrics.

Comparison of barrier height with available experiment is difficult due to the lack of experimental values for the substituted systems. For **2**, the experimentally available value of 42.27 ± 0.63 kJ/mol in CH$_2$Cl$_2$ solution[93,94] and 43.11 ±

1.26 kJ/mol in the pure liquid[95] agrees well with the calculated barrier of 43.60 kJ/mol for ε = 12.0 and 44.57 kJ/mol in water. Figure 10 shows a general trend of calculated E_a for the three systems studied. The increase in barrier to interconversion to the *trans* form shows steady increase for both **3** and **2**, corresponding to the stronger preference for the *cis* isomer with increasing dielectric, somewhat attenuated in the parent system as observed above. The sulfur system barrier is considerably less than the other two, again reflecting the relatively small increase in stabilization with increase in dielectric.

Conclusions

Detailed chemical treatments of molecular and electronic structure, including the effects of the environment, can offer

Optimization of DFT-D including effects of solvent

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2785**

significant challenges unless highly accurate methods are employed. We present here an accurate and yet cost efficient way of including solvent effects with dispersion enabled density functional theory, such that a self-consistency is maintained with respect to the solvent charges and the interaction potential. Numerical values for corresponding s_6 semiempircal dispersion parameters are reported. As shown, values of optimal s_6 parameters for solvation theory differ only slightly from that used in the vacuum, for each of the basis sets considered. Results include indirectly, consideration of BSSE. The resulting DFT-D + COSab procedure can be extended with good accuracy for prediction of interaction energies in solution.

We have applied the DFT-D + COSab method to one category of push−pull conjugated heterocycles, with known challenges associated with prediction of isomer preference and barrier to interconversion in solution environment. The behavior of changing donor/acceptor on both gas and solution phase, conformational preferences, and internal rotational barrier show the importance of dispersion, hydrogen bonding, and solvation in the computational model. The dispersion corrected DFT model with appropriate basis set reproduces the trends in the known experimental data and projects trends across a large span of dielectric.

Supporting Information Available: Table of experimental relative energies of *cis*-, *trans*-, and transition state of furfural in different solvents. This material is available free of charge via the Internet at http://pubs.acs.org.

References

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

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

(3) Hill, J. G.; Platts, J. A.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2008**, *8*, 4072–4078.

(4) Silvestrelli, P. L. *Phys. Rev. Lett.* **2008**, *100*, 053002.

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

(6) Riley, K. E.; Vondrasek, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5555–5560.

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

(8) Grimme, S.; Antony, J.; Schwabe, T.; Mück-Lichtenfeld, C. *Org. Biomol. Chem.* **2007**, *5*, 741–758.

(9) Grimme, S.; Mück-Lichtenfeld, C.; Antony, J. *J. Phys. Chem. C* **2007**, *111*, 11199–11207.

(10) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys.* **2007**, *126*, 124115.

(11) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*-154116.

(12) Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42–45.

(13) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1009–1018.

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

(15) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2006**, *28*, 555–569.

(16) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108–034115.

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

(18) von Lilienfeld, O. A.; Tavernelli, L.; Rothlisberger, U. *Phys. Rev. Lett.* **2004**, *93*, 153004.

(19) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

(20) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748.

(21) Mooij, W. T. M.; van Duijneveldt, F. B.; van Duijnevelt-van de Rijdt, J. G. C. M.; van Eijck, B. P. *J. Phys. Chem. A.* **1999**, *103*, 9872.

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

(23) Meijer, E. J.; Sprik, M. *J. Chem. Phys.* **1996**, *105*, 8684.

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

(25) Hobza, P.; Zahradnik, R. *Chem. Rev.* **1988**, *88*, 871.

(26) Kutzelnigg, W. Pair correlation theories. In *Modern Theoretical Chemistry*; Schaefer, H. F., Ed.; Plenum Press: New York, London, 1977; pp 129−188.

(27) Kutzelnigg, W. *J. Mol. Struct. (Theochem)* **1988**, *181*, 33.

(28) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; J. Wiley: New York, 2000.

(29) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(30) Bates, D. M.; Anderson, J. A.; Oloyede, P.; Tschumper, G. S. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2775–2779.

(31) Bachorz, R. A.; Bischoff, F. A.; Hoefener, S.; Klopper, W.; Ottiger, P.; Leist, R.; Frey, J. A.; Leutwyler, S. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2758–2766.

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

(33) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. *J. Phys. Chem. A* **2007**, *111*, 1146–1154.

(34) Peverati, R.; Baldridge, K. K. *J. Chem. Theory Comput.* **2008**, *4*, 2030–2048.

(35) Baldridge, K. K.; Jonas, V. *J. Chem. Phys.* **2000**, *113*, 7511.

(36) Cammi, R.; Frediani, L.; Mennucci, B.; tomasi, J.; Ruud, K.; Mikkelsen, K. V. *J. Chem. Phys.* **2002**, *117*, 13.

(37) Chen, W.; Gordon, M. S. *J. Chem. Phys.* **1996**, *105*, 11081–11090.

(38) Christiansen, O.; Mikkelsen, K. V. *J. Chem. Phys.* **1999**, *110*, 8348.

(39) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Chem. Phys.* **2001**, *114*, 5691–5701.

(40) Cramer, C. J.; Truhlar, D. G. *Continuum Solvation Models: Classical and Quantum Mechanical Implementation*; VCH Publishers: New York, 1995; Vol. 6.

(41) Giesen, D. J.; Gu, M. Z.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1996**, *61*, 8720.

(42) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

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

(44) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.

(45) Schmidt, M.; Baldridge, K. K.; Boatz, J. A.; Elbert, S.; Gordon, M.; Jenson, J. H.; Koeski, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

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

(47) Perdew, J. P. *Electronic Structure of Solids*; Akademie Verlag: Berlin, 1991.

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

(49) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

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

(51) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(52) Baldridge, K.; Klamt, A. *J. Chem. Phys.* **1997**, *106*, 6622–6633.

(53) Gregerson, L. N.; Baldridge, K. K. *Helv. Chim. Acta* **2003**, *86*, 4112–4132.

(54) Baldridge, K. K.; Jonas, V. *J. Chem. Phys.* **2000**, *113*, 7511.

(55) Klamt, A.; Jonas, V. *J. Chem. Phys.* **1996**, *105*, 9972.

(56) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(57) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, C. W. *J. Phys. Chem.* **1998**, *102*, 5074–5085.

(58) Zimmerli, U.; Parrinello, M.; Koumotsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693.

(59) Kubar, T.; Jurecka, P.; Cerny, J.; Rezáe, J.; Otyepka, M.; Valdés, H.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 5642–5647.

(60) Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* **1977**, *19*, 119.

(61) Hepburn, J.; Scoles, G. *Chem. Phys. Lett.* **1975**, *36*, 451.

(62) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(63) Klamt, A.; Eckert, F.; Hornig, M. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 355–365.

(64) Fortunelli, A.; Tomasi, J. *Chem. Phys. Lett.* **1994**, *231*, 34.

(65) Luque, F. J.; Bachs, M.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 847.

(66) Stefanovich, E. V.; Truong, T. N. *Chem. Phys. Lett.* **1995**, *244*, 65.

(67) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 877–887.

(68) Gregerson, L. N. Outlying Charge, Stability, Efficiency, and Extensibility in Quantum Mechanical Solvation Methods: the COSab-GAMESS Method, Ph.D. Thesis, University of California, San Diego, San Diego, 2004.

(69) Chiorescu, I.; Dueubel, D. V.; Arion, V. B.; Keppler, B. K. *J. Chem. Theory Comput.* **2008**, *4*, 499–506.

(70) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard III, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.

(71) Mu, W.-H.; Chasse, G. A.; Fang, D.-C. *Int. J. Quantum Chem.* **2008**, *108*, 1422–1434.

(72) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760–768.

(73) Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. J. *Theor. Chem. Acc.* **2009**, *42*, 489–492.

(74) Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2009**, *42*, 493–497.

(75) Colominas, C.; Luque, F. J.; Orozco, M. *J. Phys. Chem. A* **1999**, *103*, 6200–6208.

(76) Podolyan, Y.; Gorb, L.; Leszezynski, J. *J. Phys.Chem. A* **2002**, *106*, 12103–12109.

(77) Bende, A.; Suhai, S. *Int. J. Quantum Chem.* **2005**, *103*, 841–853.

(78) Kratochvil, M.; Engkvist, O.; Sponer, J.; Jungwirth, P.; Hobza, P. *J. Phys. Chem. A* **1998**, *102*, 6921–6926.

(79) Benassi, R.; Bertarini, C.; Hilfert, L.; Kempter, G.; Kleinpeter, E.; Spindler, J.; Taddei, F.; Thomas, S. *J. Mol. Struct. (Theochem)* **2000**, *520*, 273–294.

(80) Benassi, R.; Bertarini, C.; Kleinpeter, E.; Taddei, F.; Thomas, S. *J. Mol. Struct. (Theochem)* **2000**, *498*, 201–215.

(81) Kleinpeter, E.; Schroth, W.; Pihlaja, K. *Magn. Reson. Chem.* **2005**, *29*, 223–230.

(82) Neuvonen, H.; Fülöp, F.; Neuvonen, K.; Koch, A.; Kleinpeter, E. *J. Phys. Org. Chem.* **2008**, *21*, 173–184.

(83) Hinchliffe, A.; Mkadmh, A.; Nikolaidi, B.; Sosc'un, H. J.; Abu-Awwad, F. M. *Cent. Eur. J. Chem.* **2006**, *4*, 745–759.

(84) Puterová, Z.; Krutošíková, A.; Lyčka, A.; Ďurčeková, T. *Molecules* **2004**, *9*, 241–255.

(85) Dahlqvist, K.-I.; Forsen, S. *J. Phys. Chem.* **1965**, *69*, 4062–4071.

(86) Baldridge, K. K.; Jonas, V.; Bain, A. *J. Chem. Phys.* **2000**, *113*, 7519.

(87) Bain, A. D.; Hazendonk, P. *J. Phys. Chem. A* **1997**, *101*, 7182–7188.

(88) Little, T. S.; Qiu, J.; Durig, J. R. *Spectrochim. Acta* **1989**, *45A*, 789.

(89) Mönnig, F.; Dreizler, H.; Rudolph, H. D. *Z. Naturforsch., A: Phys Sci.* **1965**, *20*, 1323.

(90) Abraham, R. J.; Siverns, T. M. *Tetrahedron* **1972**, *28*, 3015–3023.

(91) Bertran, J. F.; Ortiz, E.; Ballester, L. *J. Mol. Struct.* **1973**, *17*, 161.

(92) Marstokk, K.-M.; Mollendal, H. *J. Mol. Struct.* **1974**, *23*, 93.

(93) Casarini, D.; Lunazzi, L.; Macciantelli, D. *J. Chem. Soc., Perkin Trans. 2* **1985**, 1839–1844.

(94) Lunazzi, L.; Placucci, G.; Chatgilialoglu, C.; Macciantelli, D. *J. Chem. Soc., Perkin Trans. 2* **1984**, 819.

(95) Pethrick, R. A.; Wyn-Jones, E. *J. Chem. Soc. A* **1969**, 713.

JCTC Journal of Chemical Theory and Computation

Structures and Energetics of SrFeO$_{2.875}$ Calculated within the GGA + *U* Framework

Wen Lai Huang* and Qingshan Zhu

State Key Laboratory of Multi-Phase Complex Systems, Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

Abstract: The energetics and electronic properties of SrFeO$_{2.875}$ have been systematically calculated with the fully relaxed atomic positions at both GGA and GGA + *U* levels, and different spin-polarized configurations have been considered. Many atoms besides the nearest neighbors of the oxygen vacancy have been found to be influenced by the vacancy in terms of positions and electronic structures. The obtained magnetic moments suggest the high-spin character of the Fe 3*d* electrons, in combination with the larger exchange splitting compared with the crystal field splitting. The local states at the Fermi level are found to be situated within the π^* band in the nonmagnetic case and the σ^* bands in the ferromagnetic and antiferromagnetic cases. The energy values concerning the oxygen-vacancy formation from SrFeO$_3$ to SrFeO$_{2.875}$ have been deduced with the correction of the O$_2$ overbinding error and the consideration of the oxygen partial pressure and the temperature.

1. Introduction

SrFeO$_{3-\delta}$ species have been investigated extensively, due to their interesting electronic, magnetic, and transport properties. Their flexible accommodation of highly mobile oxygen vacancies, in combination with high electronic conductivity, makes them appealing for applications such as electrodes of solid oxide fuel cells, electrochemical catalysts, membranes for oxygen separation, and gas sensors. At low temperatures (below certain composition-dependent transition temperatures), the oxygen-vacancy ordered phases have been observed as the Sr$_n$Fe$_n$O$_{3n-1}$ series, where n = 1, 2, 4, 8, and ∞ after updated by the report of SrFeO$_2$.[1] However, at elevated temperatures the cubic perovskite structure remains up to δ = 0.5 with disordered oxygen vacancies, and corresponding research is essential to the high-temperature applications.

Although the perfect SrFeO$_3$ has been explored theoretically at various levels, the computational investigation on the defective SrFeO$_{3-\delta}$ with $\delta \neq 0$ seems relatively limited. Koslowski[2] has analyzed the electronic structures of the cubic nonmagnetic (NM) SrFeO$_{3-\delta}$ with $0 \leq \delta \leq 0.5$, using simple but physically transparent models. It is found that with increas-

ing δ, the density of states (DOS) at the Fermi level (E_F) decreases, and the localization of eigenfunctions at E_F increases. Meanwhile, the obtained E_F lies at the top of the σ^* band, independent of δ, which has been taken as a discrepancy from the previous hypothesis of Goodenough.[3] Shein et al.[4] have calculated the electronic properties of the cubic ferromagnetic (FM) SrFeO$_{2.875}$ at the generalized gradient approximation (GGA) level based on the density functional theory (DFT) and obtained the vacancy formation energy of 2.90 eV.

The present work aims to analyze systematically the structural and electronic properties of the cubic SrFeO$_{2.875}$ at the GGA + *U* level, considering the NM, FM, and G-type antiferromagnetic (AFM) configurations. Since such a level might not be suitable for metallic phases,[5] the pure GGA level is adopted as well to facilitate comparison. In order to evaluate the influence of oxygen-vacancy formation, and to deduce the formation energy, equivalent calculations are performed on the SrFeO$_3$ counterparts. The thermodynamic corrections are provided to account for the oxygen partial pressure and temperature.

2. Computational Details

The adopted atomic positions for the NM, FM and AFM SrFeO$_3$ are illustrated in Figure 1 (a) to (c), respectively. The experimental lattice parameter a = 3.869 Å was

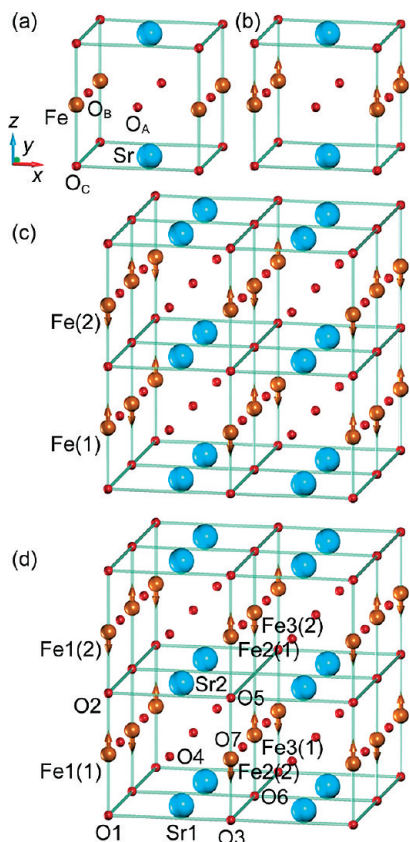* Corresponding author phone: +86-10-62650558; fax: +86-10-62536108; e-mail: wlhuang@home.ipe.ac.cn.

2788 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Huang and Zhu

Figure 1. The adopted (a) NM and (b) FM $SrFeO_3$ unit cells and AFM (c) $(SrFeO_3)_8$ and (d) unrelaxed $(SrFeO_{2.875})_8$ supercells. The arrows show the majority-spin orientations of the Fe atoms in the spin-polarized cases.

employed in the present work as in the literature.[6] As for the AFM $SrFeO_3$, we utilized a $(SrFeO_3)_8$ supercell rather than the body-centered tetragonal primitive cell $(SrFeO_3)_2$ (as adopted by Jaya et al.[6]) so as to decompose the Fe $3d$ orbitals into comparative components, and in light of the majority-spin directions, the Fe atoms are divided into Fe(1) and Fe(2) types. The $SrFeO_{2.875}$ is modeled via a $(SrFeO_{2.875})_8$ supercell, built by removing the central oxygen atom of the $(SrFeO_3)_8$ supercell and then shrinking the lattice constant from $2a$ to $2a_p$, where $a_p = 3.864$ Å as treated in the literature.[1] According to the coordination environments and distances from the oxygen vacancy, Fe, Sr, and O atoms are classified into three (Fe1 to Fe3), two (Sr1 and Sr2), and seven (O1 to O7) kinds, respectively. In the AFM $(SrFeO_{2.875})_8$ supercell, as shown in Figure 1 (d), each type of Fe atom can be distinguished further based on their majority-spin orientations, e.g., Fe1(1) or Fe1(2). To manifest the two different computational levels, the cases corresponding to the pure GGA level are labeled as NM1, FM1 and AFM1, while those at the GGA $+ U$ level are denoted by NM2, FM2, and AFM2.

Calculations were conducted via the plane-wave pseudo-potential (PW-PP) DFT method as implemented in the Quantum-ESPRESSO package.[7] The Perdew-Burke-Ernzerhof (PBE)[8] GGA was employed for the exchange-correlation functional. Ultrasoft pseudopotentials (US-PPs) were used, where $3d$ and $4s$ states were considered for Fe with 8 e, $4s$, $4p$, $4d$, $5s$, and $5p$ states for Sr with 10 e, and $2s$ and $2p$ states for O with 6 e. All the pseudopotentials were generated

Table 1. Displacements (Å) of the Atoms in $SrFeO_{2.875}$ Due to Relaxation

	NM1	NM2	FM1	FM2	AFM1	AFM2
Fe1	0.015	0.004	0.013	0.015	0.012	0.008
Fe2	0.022	0.020	0.023	0.022	0.021	0.016
Fe3	0.191	0.204	0.080	0.048	0.074	0.058
Sr1	0.101	0.101	0.040	0.031	0.028	0.036
Sr2	0.061	0.072	0.093	0.093	0.060	0.070
O1	0.000	0.000	0.000	0.000	≈ 0.000	≈ 0.000
O2	0.000	0.000	0.000	0.000	≈ 0.000	≈ 0.000
O3	0.000	0.000	0.000	0.000	≈ 0.000	≈ 0.000
O4	0.031	0.023	0.011	0.017	0.029	0.034
O5	0.000	0.000	0.000	0.000	≈ 0.000	≈ 0.000
O6	0.000	0.000	0.000	0.000	≈ 0.000	≈ 0.000
O7	0.180	0.183	0.144	0.174	0.209	0.218

with scalar-relativistic calculations, and nonlinear core corrections (NLCC) were included in the pseudopotentials of Fe and Sr. Within the GGA $+ U$ scheme, the value of $U = 4.3$ eV was applied to the Fe species.

The kinetic energy cutoffs of 40 and 500 Ry were employed for the smooth part of the electronic wave functions and the augmented electron density respectively to achieve a total-energy difference below 1 meV/atom, accompanied by a Monkhorst-Pack[9] k-point grid (centered at the Γ point) of $8 \times 8 \times 8$ for $SrFeO_3$ and $4 \times 4 \times 4$ for $(SrFeO_3)_8$ and $(SrFeO_{2.875})_8$. The cold smearing[10] of 0.01 Ry was selected, and all the atomic positions in $(SrFeO_{2.875})_8$ were relaxed until all components of the residual forces were smaller than 0.026 eV/Å via the Broyden, Fletcher, Goldfarb, Shanno (BFGS)-based method.[11] The tetrahedron method[12] was exploited to calculate the DOS, and the atomic populations were analyzed using both the Löwdin[13] (included in the Quantum-ESPRESSO package[7]) and the Bader[14] (by means of the program developed by Henkelman et al.[15]) methods.

In order to figure out the oxygen-vacancy formation energy during the transformation from $SrFeO_3$ to $SrFeO_{2.875}$, the triplet O_2 was relaxed, and subsequent calculations were performed in a 10 Å cubic box, as employed by Choi et al.[16] with a k point at the Γ. The kinetic energy cutoffs of 80 and 500 Ry were chosen for wave functions and charge density, respectively, to ensure a total-energy difference below 2 meV/atom.

To determine the reduction limit of the O_2 partial pressure, the total energy values of Fe and SrO were computed additionally with the experimental lattice parameters,[17] using the same cutoffs as for $SrFeO_3$, and the k points were sampled with a separation about 0.04 1/Å in each of the three reciprocal lattice directions.

3. Results and Discussion

3.1. Atomic Positions and Populations. As explicated in Table 1, the relaxation generally leads to apparent changes in the positions of the nearest-neighbor (NN) (Fe3, Sr2, and O7) atoms of the oxygen vacancy in $SrFeO_{2.875}$, but many atoms beyond experience certain influences as well, especially Sr1 atoms whose displacements (away from the vacancy equally along the x- and y-axes, and additional variation also appears along the z-axis for the AFM cases)

are even more sizable than Sr2 in the NM cases. All the Fe1 and Fe2 atoms move along the z-axis. The Fe2 atoms in all cases and the Fe1 atoms in the NM cases approach the vacancy, while the Fe1 atoms in the FM and AFM cases are displaced away from the vacancy. The O4 atoms shift away from (or toward) the vacancy along the x-axis in the FM (or NM and AFM) cases and away from the vacancy along the z-axis in all the cases, and the displacements along the z-axis are more significant than along the x-axis. In comparison with the FM results calculated using the atomic orbitals as basis sets,[4] the present work offers the same trend and comparable displacement values for the NN atoms. The eight NN O7 atoms relax toward the vacancy (basically along the z-axis), whereas the two NN Fe3 and four NN Sr2 atoms retreat from the vacancy (along the z-axis for the former and equally along the x-axis and y-axis for the latter except for the AFM cases where slight shifts along the z-axis have also occurred). Such results can be interpreted based mostly on the electrostatic interactions within the ionic pictures. According to the simple empirical model in the Kröger-Vink notation, $O \rightarrow V_O^{\bullet\bullet} + 1/2\ O_2\ (g) + 2\ e'$, the oxygen vacancy is positively charged and thus attracts the negative oxygen and repulses positive Fe and Sr ions in the lattice. Similar phenomena have been encountered in the literature, e.g., the oxygen vacancy repulses the neighboring Ti atoms, and the Sr or Ti vacancy repels the neighboring oxygen atoms in the defective SrTiO₃.[18] The absolute values of the displacements are in the order of O7 > Sr2 ≈ Fe3 for the FM and AFM configurations and of Fe3 > O7 > Sr2 in the NM cases. These findings can be roughly correlated with the corresponding charges and interatomic distances. Neglecting the charge transfer induced by the vacancy formation, Fe3 exhibits the highest nominal valence (+4), along with the shortest distance from the vacant site ($1/2\ a_p$, compared with $\sqrt{2}/2\ a_p$ for those of Sr2 or O7 before relaxation), and the largest displacements can be expected. Reduction due to the charge transfer (described below) weakens the electrostatic interaction between Fe3 or Sr2 and the vacancy and yields their smaller displacements, while reverse results might be anticipated for the negatively charged O7.

To elucidate the vacancy-incurred charge transfer, we analyzed the atomic populations in both SrFeO₃ and SrFeO₂.₈₇₅. The Löwdin results of SrFeO₃ and SrFeO₂.₈₇₅ are provided in Table 2. It is obvious that the vacancy transfers charges principally to the NN atoms, and the order of transfer amount is generally Fe3 > Sr2 ≈ O7. To make a crosscheck on the tendency, the Bader analyses were complemented, as in Table 3, and coherent tendencies can be observed.

A comparable amount of charge transfer to Fe3 shortens its displacement in the NM cases less than in the FM and AFM cases, owing to the relatively larger dispersivity of the 3d orbitals (that accept the charges) in the NM cases, and the Fe3 shows still greater displacements than O7 and Sr2 in the NM cases. The longer displacements of O7 than Sr2 might be ascribed qualitatively to the larger charge transfer to Sr2 as well as the steric factor: the smaller size of oxygen and more space and thus more feasibility to approach the vacancy.

Table 2. Löwdin Populations (e) of the Atoms in SrFeO₃ and SrFeO₂.₈₇₅

		NM1	NM2	FM1	FM2	AFM1	AFM2
a	Fe	6.71	6.81	6.45	6.31	6.45	6.36
	Sr	9.86	9.85	9.90	9.93	9.90	9.92
	O	6.45	6.42	6.51	6.55	6.51	6.54
b	Fe1	6.68	6.81	6.47	6.31	6.43	6.35
	Fe2	6.72	6.82	6.45	6.28	6.42	6.33
	Fe3	6.85	6.87	6.52	6.38	6.53	6.47
	Sr1	9.84	9.83	9.88	9.90	9.88	9.90
	Sr2	9.88	9.88	9.93	9.96	9.93	9.95
	O1	6.44	6.44	6.51	6.55	6.51	6.52
	O2	6.42	6.44	6.51	6.55	6.51	6.52
	O3	6.46	6.45	6.54	6.59	6.55	6.58
	O4	6.44	6.42	6.52	6.56	6.53	6.54
	O5	6.44	6.41	6.52	6.55	6.52	6.55
	O6	6.40	6.33	6.50	6.63	6.55	6.58
	O7	6.48	6.45	6.54	6.57	6.54	6.56

a SrFeO₃. *b* SrFeO₂.₈₇₅.

Table 3. Bader Populations (e) of the Atoms in SrFeO₃ and SrFeO₂.₈₇₅

		NM1	NM2	FM1	FM2	AFM1	AFM2
a	Fe	6.21	6.29	6.09	6.02	6.08	6.05
	Sr	8.37	8.37	8.38	8.39	8.38	8.39
	O	7.14	7.12	7.18	7.20	7.18	7.19
b	Fe1	6.21	6.29	6.09	5.99	6.04	6.09
	Fe2	6.22	6.31	6.13	6.03	6.09	6.05
	Fe3	6.39	6.37	6.25	6.19	6.21	6.21
	Sr1	8.41	8.41	8.39	8.40	8.40	8.41
	Sr2	8.41	8.41	8.42	8.42	8.42	8.42
	O1	7.19	7.19	7.19	7.23	7.20	7.17
	O2	7.20	7.18	7.22	7.25	7.21	7.24
	O3	7.17	7.16	7.22	7.24	7.23	7.25
	O4	7.16	7.14	7.22	7.24	7.21	7.21
	O5	7.17	7.12	7.22	7.21	7.23	7.20
	O6	7.03	7.06	7.13	7.21	7.18	7.17
	O7	7.15	7.13	7.18	7.21	7.21	7.21

a SrFeO₃. *b* SrFeO₂.₈₇₅.

Introduction of the Hubbard U term enlarges the displacements of O7 in all cases, Fe3 in the NM case, and Sr2 in the NM and AFM cases, while it reduces those of Fe3 in the FM and AFM cases. Electrostatic analysis can offer a reasonable explanation again. Take the displacements of Fe3 as an example. In comparison with the GGA method, the GGA + U level generates smaller occupancy difference between Fe1 and Fe3 in the NM case and larger ones in the FM and AFM cases. Assuming that the Fe1 atoms receive no influence and charge transfer from the vacancy formation, the above results mean that the charge transfer to Fe3 is smaller in the NM case, and larger in the FM and AFM cases, when the U term is introduced. This is in accord with the displacement changes.

To describe the charge distribution among the Fe 3d orbitals, the relevant Löwdin populations are listed in Table 4. Referring to SrFeO₃, the Fe3 atoms in NM SrFeO₂.₈₇₅ accept electrons largely on the $3d_{z^2}$ or $3d_{x^2-y^2}$ orbital. It is also evident that the $3d_{xz}$, $3d_{yz}$, and $3d_{xy}$ orbitals contribute more than the $3d_{z^2}$ and $3d_{x^2-y^2}$ ones to the Fe magnetic moments in the FM and AFM cases. The oxygen vacancy enhances the spin polarization on the Fe3 $3d_{z^2}$ orbital and weakens that on the Fe3 $3d_{x^2-y^2}$ orbital, while the Fe2 atoms

Table 4. Löwdin Populations (e) of the Fe or Fe(1) 3*d* Orbitals in Different Cases

	z^2		xz		yz		$x^2 - y^2$		xy		total	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
NM1[a]	0.95		1.52		1.52		0.95		1.52		6.45	
NM2[a]	0.79		1.66		1.66		0.79		1.66		6.55	
FM1[a]	0.77	0.34	1.00	0.32	1.00	0.32	0.77	0.34	1.00	0.32	4.53	1.65
FM2[a]	0.88	0.36	0.99	0.20	0.99	0.20	0.88	0.36	0.99	0.20	4.73	1.31
AFM1[b]	0.78	0.46	0.98	0.26	0.98	0.26	0.78	0.46	0.98	0.26	4.50	1.68
AFM2[b]	0.86	0.46	0.99	0.16	0.99	0.16	0.86	0.46	0.99	0.16	4.68	1.41
NM1[c]	0.92		1.54		1.54		0.91		1.52		6.43	
NM1[d]	0.92		1.50		1.58		0.94		1.53		6.47	
NM1[e]	1.12		1.49		1.49		0.96		1.54		6.60	
NM2[c]	0.75		1.70		1.70		0.74		1.68		6.56	
NM2[d]	0.74		1.60		1.79		0.76		1.68		6.57	
NM2[e]	0.78		1.66		1.66		0.88		1.63		6.61	
FM1[c]	0.79	0.34	0.99	0.32	0.99	0.32	0.77	0.35	0.99	0.34	4.54	1.67
FM1[d]	0.81	0.32	0.99	0.29	0.99	0.32	0.86	0.31	0.99	0.29	4.65	1.53
FM1[e]	0.86	0.31	0.99	0.35	0.99	0.35	0.70	0.39	0.99	0.32	4.53	1.72
FM2[c]	0.89	0.35	0.99	0.19	0.99	0.19	0.88	0.36	0.99	0.19	4.75	1.28
FM2[d]	0.90	0.34	0.99	0.19	0.99	0.15	0.93	0.33	0.99	0.18	4.81	1.19
FM2[e]	0.95	0.29	0.99	0.22	0.99	0.22	0.81	0.41	0.99	0.22	4.74	1.36
AFM1[f]	0.79	0.45	0.99	0.24	0.99	0.24	0.80	0.44	0.99	0.24	4.55	1.61
AFM1[g]	0.81	0.41	0.99	0.24	0.99	0.23	0.88	0.38	0.99	0.23	4.66	1.50
AFM1[h]	0.92	0.41	0.98	0.31	0.98	0.31	0.58	0.49	0.99	0.28	4.46	1.80
AFM2[f]	0.86	0.46	0.99	0.16	0.99	0.16	0.87	0.46	0.99	0.16	4.69	1.39
AFM2[g]	0.89	0.43	0.99	0.16	0.99	0.14	0.93	0.38	0.99	0.15	4.80	1.26
AFM2[h]	0.95	0.42	0.99	0.22	0.99	0.22	0.68	0.51	0.99	0.21	4.60	1.59

[a] Fe in $SrFeO_3$. [b] Fe(1) in $SrFeO_3$. [c] Fe1 in $SrFeO_{2.875}$. [d] Fe2 in $SrFeO_{2.875}$. [e] Fe3 in $SrFeO_{2.875}$. [f] Fe1(1) in $SrFeO_{2.875}$. [g] Fe2(1) in $SrFeO_{2.875}$. [h] Fe3(1) in $SrFeO_{2.875}$.

show a vacancy-induced increase in the magnetic moments of almost all five 3*d* orbitals.

3.2. Electronic Band Structures. Regardless of the reduction from $SrFeO_3$ to $SrFeO_{2.875}$, the magnetic patterns (NM, FM, or AFM), and the change in computational level from pure GGA to GGA + *U*, the band sequence remains unchanged: Sr 4*s*, O 2*s*, and Sr 4*p* form the lowest three occupied bands sequentially with increasing energy, the hybridization of O 2*p* and Fe 3*d* exists around E_F, and Sr 4*d*, Fe 3*s*, Sr 5*s*, and Sr 5*p* contribute to the bands of higher energy. Figures 2−4 depict the total and partial density of states (PDOS) of the species in NM, FM, and AFM configurations, respectively, calculated within the GGA + *U* scheme. In the displayed energy range, the Fe 4*s* states exhibit negligible density though plotted, and the trivial contribution of Sr 5*s* and 5*p* has been omitted. The (P)DOS peaks in $SrFeO_{2.875}$ are generally lower but wider than their counterparts in $SrFeO_3$, and some peaks have decomposed, due to the lowering of the symmetry.

Based on simple models, Koslowski has reported the decreasing DOS at E_F with increasing δ for NM $SrFeO_{3-\delta}$. This trend has not been reproduced in our NM1 cases (8.77 and 9.02 states/(eV fu) for $SrFeO_3$ and $SrFeO_{2.875}$, respectively) but appears somewhat in the NM2 cases (7.63 and 7.56 states/(eV fu) for $SrFeO_3$ and $SrFeO_{2.875}$, respectively), considering $\delta = 0$ and 0.125. By the way, Jaya et al. has given a value of 11.94 states/(eV fu) for NM $SrFeO_3$.[6]

The obtained values of the total magnetic moment are 3.61 and 4.05 μ_B/fu (formula) for FM1 and FM2 $SrFeO_3$, respectively, whereas the counterparts for $SrFeO_{2.875}$ are 3.79 and 4.25 μ_B/fu correspondingly. The absolute magnetic moment of AFM1 $SrFeO_3$ is 3.13 μ_B/fu, and that of AFM2 $SrFeO_3$ is 3.56 μ_B/fu. In contrast, the AFM1 and AFM2 cases of $SrFeO_{2.875}$ yield 3.36 and 3.70 μ_B/fu, respectively. It is
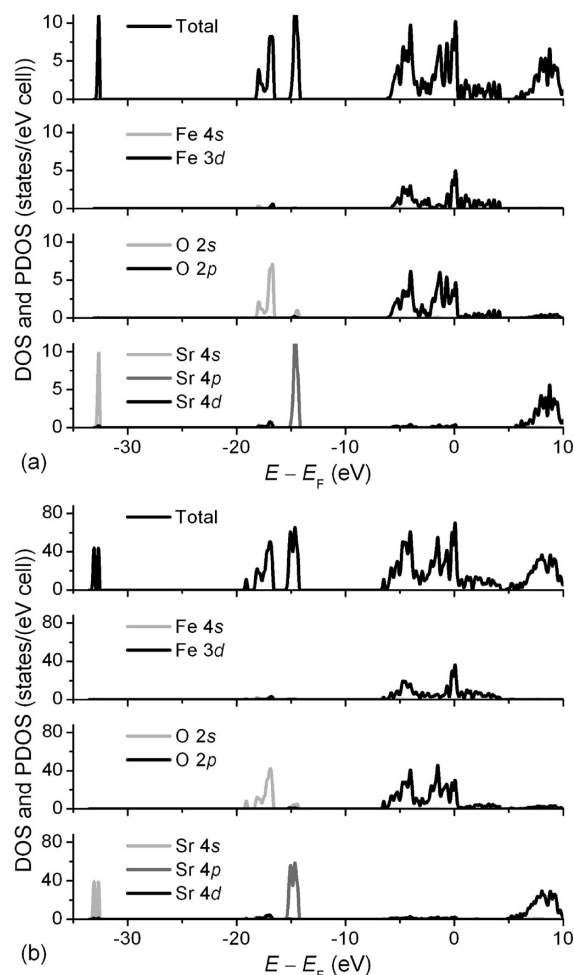
Figure 2. Total and partial DOS of NM2 (a) $SrFeO_3$ and (b) $(SrFeO_{2.875})_8$.

Structures and Energetics of SrFeO$_{2.875}$

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2791**
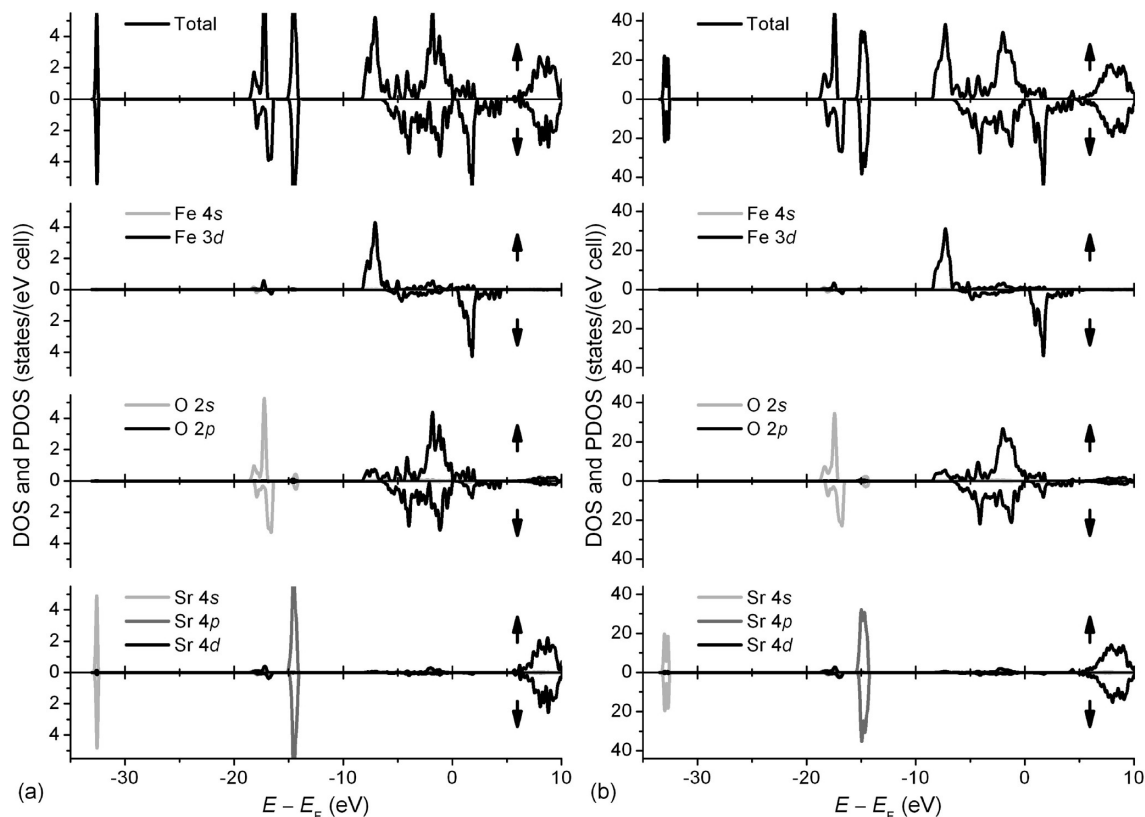
Figure 3. Total and partial DOS of FM2 (a) SrFeO$_3$ and (b) (SrFeO$_{2.875}$)$_8$.
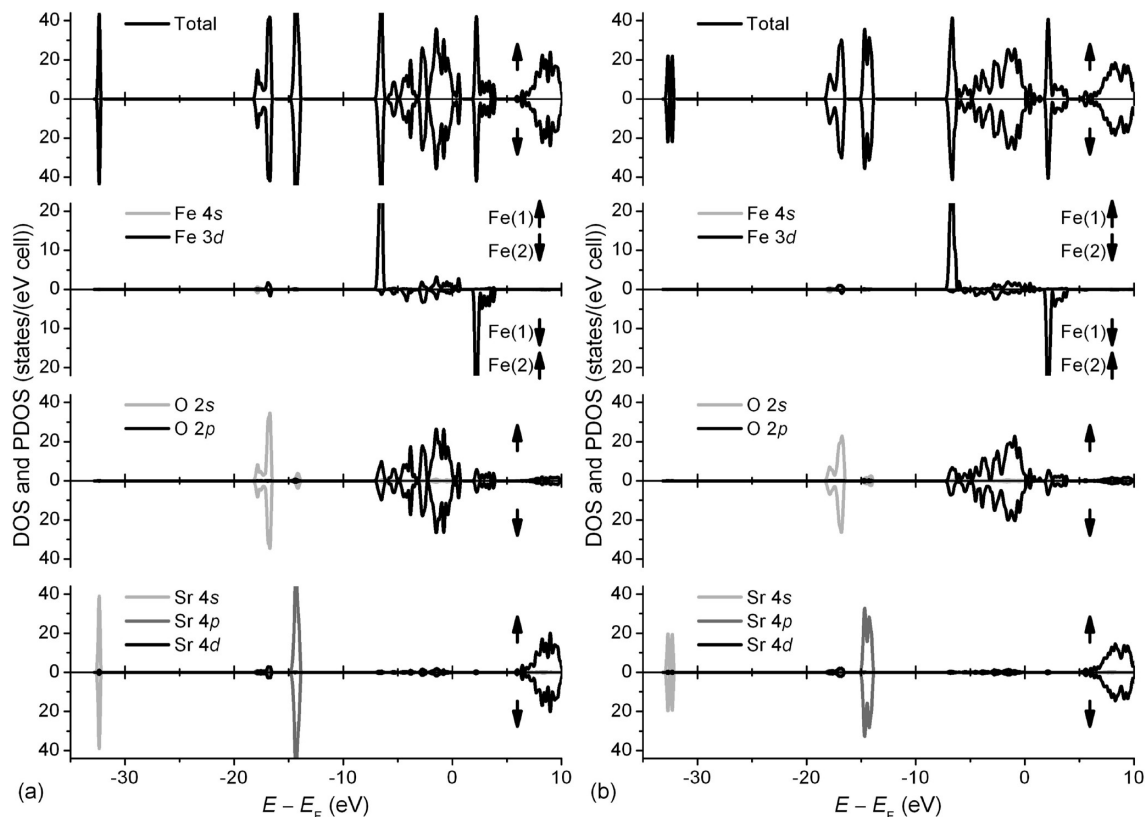
Figure 4. Total and partial DOS of AFM2 (a) (SrFeO$_3$)$_8$ and (b) (SrFeO$_{2.875}$)$_8$.

obvious that both the Hubbard U term and the oxygen vacancy have improved the magnetic moments.

In order to reveal the bonding characteristics, we provide the components of O $2p$ and Fe $3d$ orbitals and their overlapping. The GGA + U results for SrFeO$_3$ are given in

Figure 5. For the NM2 case, the σ bonding is within $-5.7 \sim -3.6$ eV, the π bonding covers $-5.6 \sim -2.3$ eV, the π^* antibonding emerges in the energy range of $-0.8-0.3$ eV, and the σ^* antibonding is apparent in 0.4–4.3 eV. The observed crystal field splitting is about 1.2 eV, and the Fermi
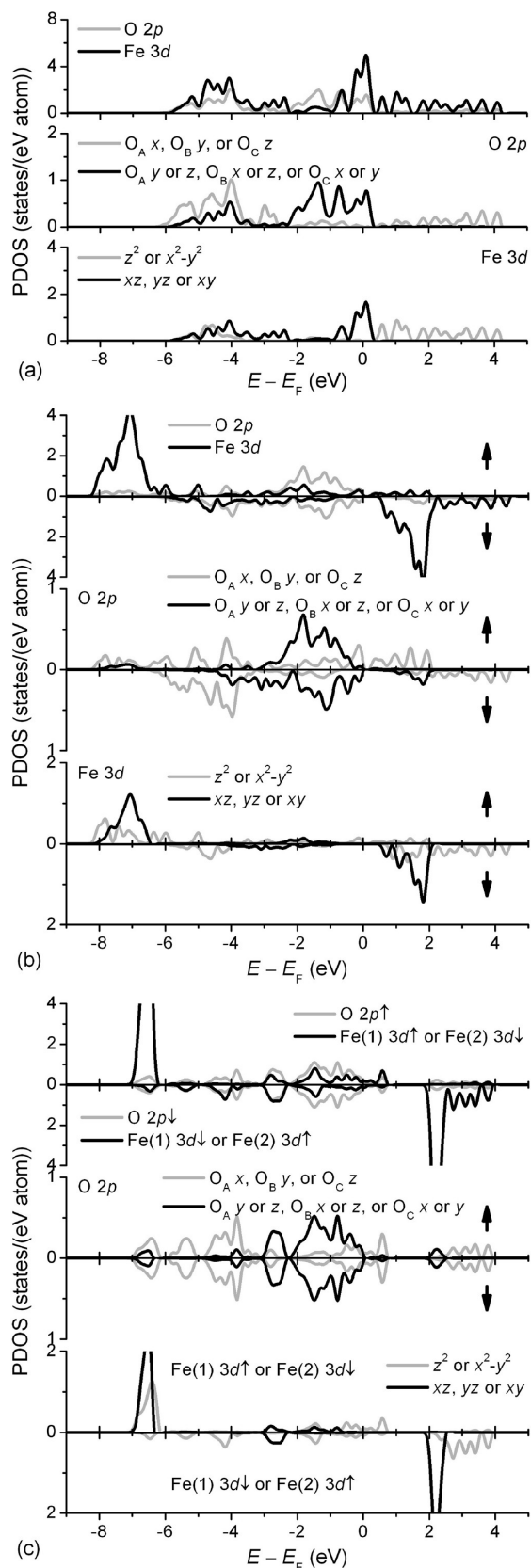
Figure 5. Overlap between O 2p and Fe 3d orbitals in (a) NM2, (b) FM2, and (c) AFM2 SrFeO₃.

Figure 6. The components of (a) O 2p and (b) Fe 3d PDOS in NM2 SrFeO₂.₈₇₅.

spin channel, the observable regions are -6.0 to -4.0, -4.4 to -0.8, 0.9–2.1, and 0.7–4.5 eV correspondingly, with the exchange splitting around 2.9 eV. The larger exchange splitting than the crystal field splitting favors the high-spin (HS) structure. Regarding the AFM2 case, the energy ranges concerning the σ, π bonding and π^*, σ^* antibonding are -7.0 to -6.2, -6.8 to -6.4, -3.0 to -0.7, and -1.8–0.7 eV, respectively, for the Fe(1) majority spin part and -6.7 to -4.0, -3.0 to -2.3, 2.0–2.5, and 2.2–4.0 eV for the Fe(1) minority spin one.

Figures 6–8 plot the PDOS of O 2p and Fe 3d orbitals in $SrFeO_{2.875}$ obtained within the GGA + U scheme. The oxygen vacancy has tailored the structural and electronic properties apparently, removing the orbital degeneracies. Referring to those in $SrFeO_3$, the 2p (as well as 2s, not shown) orbitals of O6 and O7 exhibit the most significant difference. The σ bonding and σ^* antibonding between O6 2p and Fe3 3d becomes highly localized, and in the NM2 case such bonding might be the strongest among the O 2p–Fe 3d pairs, appearing within the lowest energy range. A half-metallic feature can be noticed in FM2 $SrFeO_{2.875}$.

To further discern the states at E_F, the local DOS (LDOS) isosurfaces are presented in Figure 9, rendered via the XCrySDen code.[20] Associated with Figures 6–8, Figure 9 shows that the states at E_F correspond to the π^* antibonding

level is situated in the π^* antibonding region, which is coincident with the result of Matar.[19] In the FM2 case, the majority spin channel shows the σ, π bonding and π^*, σ^* antibonding within -8.2 to -5.8, -8.0 to -6.5, -2.5 to -1.2, and -0.3–2.1 eV, respectively, and for the minority
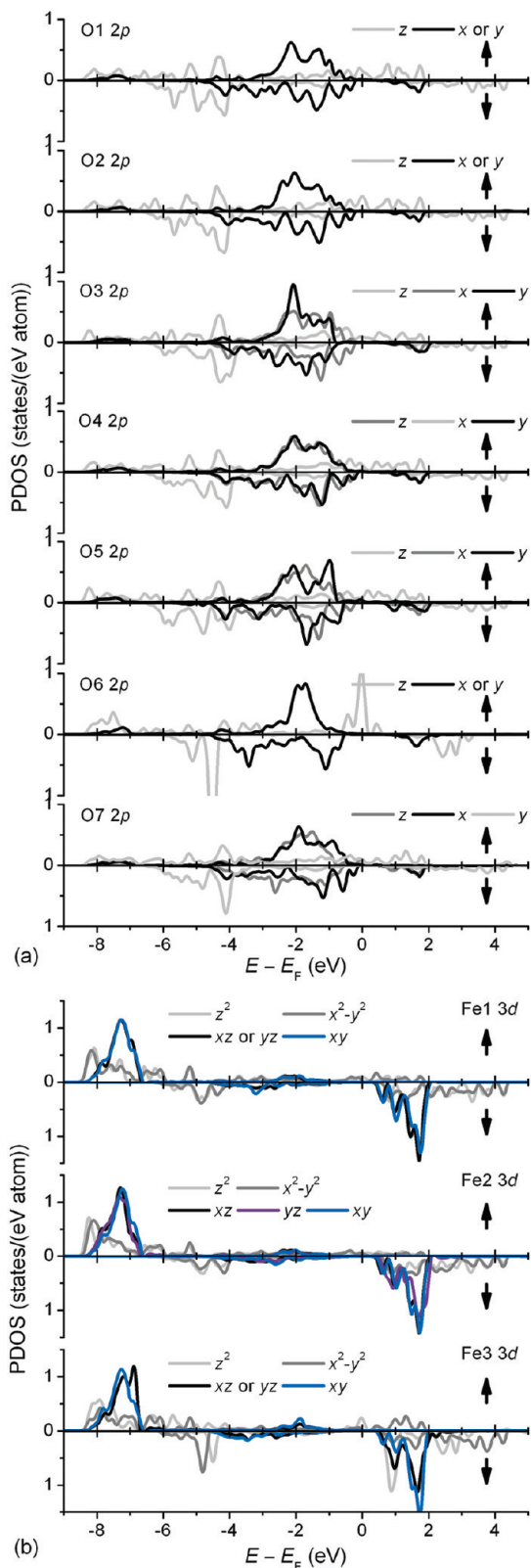
Figure 7. The components of (a) O 2p and (b) Fe 3d PDOS in FM2 SrFeO$_{2.875}$.

Figure 8. The components of (a) O 2p and (b) Fe 3d PDOS in AFM2 SrFeO$_{2.875}$.

for NM2 SrFeO$_3$ and SrFeO$_{2.875}$, and the σ^* antibonding in the spin-up channel for FM2 SrFeO$_3$ and SrFeO$_{2.875}$. The AFM2 SrFeO$_{2.875}$ displays the σ^* antibonding of the Fe1 and Fe2 3d orbitals with the 2p orbitals of adjacent oxygen atoms at E_F, while the local states of AFM2 SrFeO$_3$ at E_F manifest the nonbonding nature. Koslowski[2] has explored NM Sr-

FeO$_{3-\delta}$ using simple models and found that the σ^* antibonding states arise at E_F, which is reckoned as a difference from Goodenough's hypothesis.[3] However, Goodenough has predicted the π^* antibonding states at E_F based on low-spin (LS) Fe^{4+}, expecting FM coupling. Concerning the NM pattern, the exchange splitting is neglected, and it is
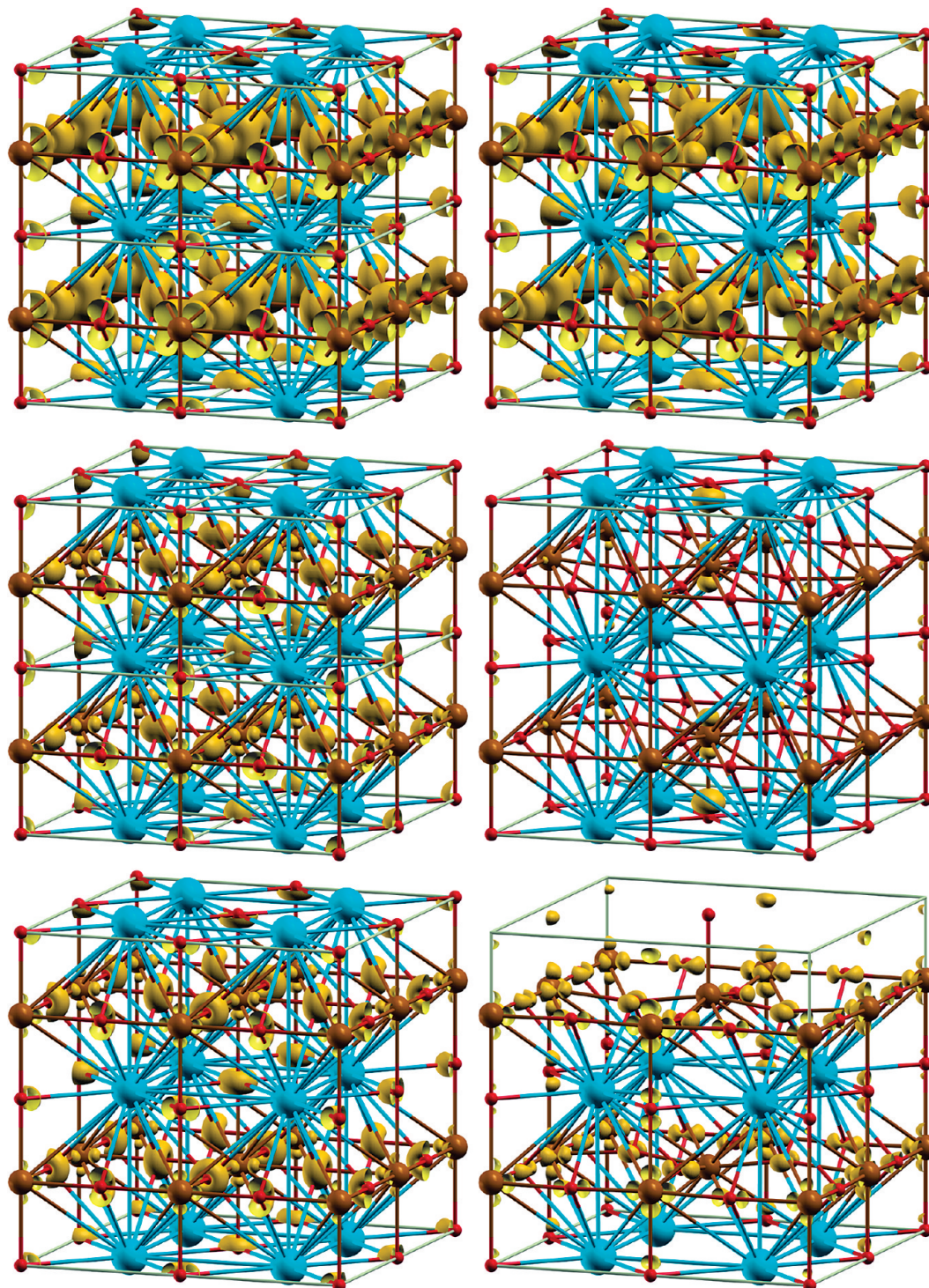
Figure 9. The local DOS isosurfaces at E_F of NM2 (top), FM2 (middle), and AFM2 (bottom) SrFeO$_3$ (left) and SrFeO$_{2.875}$ (right), rendered with an isovalue of 1.0.

reasonable that the crystal field makes the partially filled π^* orbitals and empty σ^* ones for Fe^{4+} (3d^4). This is consistent with the results of Matar[19] and Jaya et al.[6] for SrFeO$_3$, and the partially filled σ^* band achieved by Koslowski might be ascribed to some deficiencies of the Anderson-Mott-Hubbard model or its approximation.[2] Once the spin polarization is incorporated, the HS configuration has been found favorable experimentally[19] and theoretically,[6,19] which offers the σ^* states at E_F. Based on the competitive crystal field splitting and exchange splitting, and the resultant magnetic moments,

the principal HS pattern is confirmed for both SrFeO$_3$ and SrFeO$_{2.875}$ in the present work.

On the basis of the theorem of Anderson,[21] the orbital localization is expectable in disordered systems. Koslowski[2] has discovered the increasing localization of states at E_F with δ in SrFeO$_{3-\delta}$. Similarly, as depicted in Figure 9, a reduced number of states appear in SrFeO$_{2.875}$ in comparison with SrFeO$_3$, signifying the vacancy-induced localization at E_F. The orbital localization might also be implicated in the larger magnetic moments of SrFeO$_{2.875}$ configurations than those of SrFeO$_3$

counterparts, for localized electrons might carry spontaneous atomic moments while collective electrons might not.[3]

3.3. Vacancy Formation Energetics. Congruent with the reports in the literature,[19,22] both the present GGA and GGA + U calculations fail to predict the experimentally identified AFM ground state of SrFeO$_3$. At the GGA level, the FM SrFeO$_3$ was computed to be 1.308 and 0.235 eV/fu more stable than the NM and AFM patterns, respectively, while the values are 2.098 and 0.196 eV/fu with the GGA + U approach, respectively. In comparison, Matar[19] reported that the FM state is 0.224 eV/fu more stable than the AFM state, via both the LMTO-ASA (linear muffin tin orbital-atomic sphere approximation) and the ASW (augmented spherical wave) methods (where), and Shein et al.[22] gave a value of 0.262 eV/fu using the LSDA (local spin density approximation) + U formalism with U = 6.0 eV and the exchange parameter J = 0.6 eV.

For the triplet O$_2$, we obtained the relaxed O–O bond length of 1.236 Å, and the calculated vibrational frequency of 1572.2 cm^{-1}. These agree well with the results calculated under the GGA-PAW (projector augmented wave) scheme and the experimental values.[16]

Following the reduction reaction, 8 SrFeO$_3$ → (SrFeO$_{2.875}$)$_8$ + 1/2 O$_2$ (g), at the GGA level, the oxygen-vacancy formation energy is evaluated as

$$E_{\text{vf1}} = E((\text{SrFeO}_{2.875})_8)^{\text{GGA}} + 1/2 E(\text{O}_2)^{\text{GGA}} - 8E(\text{SrFeO}_3)^{\text{GGA}} \quad (1)$$

where $E(\text{O}_2)^{\text{GGA}}$ is the total energy of the triplet O$_2$, while $E((\text{SrFeO}_{2.875})_8)^{\text{GGA}}$ and $E(\text{SrFeO}_3)^{\text{GGA}}$ are the total energy values of (SrFeO$_{2.875}$)$_8$ and SrFeO$_3$ at this level, respectively. We obtain the E_{vf1} values of 0.35, 1.51, and 0.92 eV for the AFM, FM, and NM cases correspondingly. By comparison, Shein et al.[4] have reported that a value of 2.90 eV for the FM case, using the GGA approach as well, but Troullier-Martins norm-conserving pseudopotentials, the atomic orbitals basis sets, Sr 5$s^{24}p^6 4d^0$ configuration (8 *e*), and a value of 3.7054 Å for both *a* in SrFeO$_3$ and a_{p} in (SrFeO$_{2.875}$)$_8$.

It is known that the O$_2$ molecule is overbound within both the LDA and GGA frameworks (the experimental binding energy of O$_2$ is about − 5.23 eV,[23] in comparison with the GGA value[5,24] around − 5.99 or − 6.02 eV), which might cause an underestimate for the O-vacancy formation energy. By fitting the calculated formation energy of nontransition metal oxides from metals and O$_2$, and comparing with the experimental formation enthalpy values, Wang et al.[5] derived a shift of 1.36 eV/O$_2$ (i.e., 0.68 eV per oxygen atom). This is larger than the binding energy difference and is surmised to be related with the electron transferring to (oxidation) or from (reduction) the oxygen 2*p* orbitals. Therefore, the correction in this aspect might suggest an addition of 0.68 eV to the calculated O-vacancy formation energy. In the other aspect, treating 3*d* metal states using LDA or GGA causes the well-known self-interaction error that incorrectly increases the energy of the reduced state (with more d electrons), thus overestimating the O-vacancy formation energy. It is found that the GGA + U scheme can remove this error for many systems to some extent via suitable U

values.[5] Since this error is opposite to and hence cancels the O$_2$ overbinding error to a certain degree, giving rise to better agreement with experiments, it is better to correct or uncorrect these two errors simultaneously. However, the above-mentioned cancellation is rather arbitrary, and the uncorrected results are essentially unreliable. Therefore, we evaluated further the corrected E_{vf2} values, using the GGA + U energetics and the 0.68 eV/O correction as expressed in

$$E_{\text{vf2}} = E((\text{SrFeO}_{2.875})_8)^{\text{GGA}+U} + 1/2 E(\text{O}_2)^{\text{GGA}} - 8E(\text{SrFeO}_3)^{\text{GGA}+U} + 0.68 \text{ eV} \quad (2)$$

where $E((\text{SrFeO}_{2.875})_8)^{\text{GGA}+U}$ and $E(\text{SrFeO}_3)^{\text{GGA}+U}$ are the total energy values of (SrFeO$_{2.875}$)$_8$ and SrFeO$_3$ achieved at the GGA + U level. The E_{vf2} results are 0.23, 1.35, and 1.32 eV for the AFM, FM, and NM structures, respectively.

Neglecting the *pV* (pressure *p* and the cell volume *V*) and vibrational contributions of the solid phases as in the literature,[25−27] the temperature and pressure dependence of $E_{\text{vf}i}$ (*i* = 1 and 2) can be deduced simply by replacing the 1/2 $E(\text{O}_2)^{\text{GGA}}$ term in eqs 1 and 2 with the O chemical potential $\mu_\text{O}(T, p(\text{O}_2))$ at the temperature *T* and the O$_2$ partial pressure $p(\text{O}_2)$, and the resultant expressions are

$$E_{\text{vf1}}(T, p(\text{O}_2)) = E((\text{SrFeO}_{2.875})_8)^{\text{GGA}} + \mu_\text{O}(T, p(\text{O}_2)) - 8E(\text{SrFeO}_3)^{\text{GGA}} \quad (3)$$

$$E_{\text{vf2}}(T, p(\text{O}_2)) = E((\text{SrFeO}_{2.875})_8)^{\text{GGA}+U} + \mu_\text{O}(T, p(\text{O}_2)) - 8E(\text{SrFeO}_3)^{\text{GGA}+U} + 0.68 \text{ eV} \quad (4)$$

where $\mu_\text{O}(T, p(\text{O}_2)) = 1/2\, E(\text{O}_2)^{\text{GGA}} + 1/2\, \Delta G(\Delta T, p^\text{o}(\text{O}_2), \text{O}_2) + 1/2\, k_\text{B}T \ln(p(\text{O}_2)/p^\text{o}(\text{O}_2))$, k_B is the Boltzmann constant, and $p^\text{o}(\text{O}_2)$ is the O$_2$ partial pressure at the standard state, i.e., 1 atm. Using the 1/2 $\Delta G(\Delta T, p^\text{o}(\text{O}_2), \text{O}_2)$ values tabulated by Reuter and Scheffler,[25] the $E_{\text{vf}i}(T, 1\text{ atm}) \sim T$ and $E_{\text{vf}i}(700$ K, $p(\text{O}_2)) \sim p(\text{O}_2)$ plots are provided in Figure 10. Different magnetic configurations have been presented here to facilitate mutual comparison; however, it should be noted that the T_N of SrFeO$_3$ is around 130 or 134 K,[6,19] and in ordered SrFeO$_{2.875}$ the magnetic pattern was observed merely below about 85 K.[28] The lower and upper limits of $p(\text{O}_2)$ in Figure 10(b) were determined using the technique analogous to that in the literature for other systems.[25,27] The O-rich limit is defined as the point beyond which gas phase O starts to condense,[25,26] that is, let $\mu_\text{O}(T, p(\text{O}_2)^\text{max}) = 1/2\, E(\text{O}_2)^{\text{GGA}} + 1/2\, \Delta G(\Delta T, p^\text{o}(\text{O}_2), \text{O}_2) + 1/2\, k_\text{B}T \ln(p(\text{O}_2)^\text{max}/p^\text{o}(\text{O}_2)) = 1/2\, E(\text{O}_2)^{\text{GGA}}$, and obtain $p(\text{O}_2)^\text{max} = 3.25 \times 10^{10}$ atm at $T = 700$ K. The O-poor limit is determined considering the coexistence of NM SrFeO$_3$ (the FM and AFM cases give much lower limits), the body-centered cubic (bcc) FM Fe, SrO, and gaseous triplet O$_2$. Within the GGA scheme, let $\mu_\text{O}(T, p(\text{O}_2)^{\text{GGA,min}}) = 1/2\, E(\text{O}_2)^{\text{GGA}} + 1/2\, \Delta G(\Delta T, p^\text{o}(\text{O}_2), \text{O}_2) + 1/2\, k_\text{B}T \ln(p(\text{O}_2)^{\text{GGA,min}}/p^\text{o}(\text{O}_2)) = 1/2\, E(\text{SrFeO}_3)^{\text{GGA}} - 1/2\, E(\text{Fe})^{\text{GGA}} - 1/2\, E(\text{SrO})^{\text{GGA}}$, where $E(\text{SrFeO}_3)^{\text{GGA}}$, $E(\text{Fe})^{\text{GGA}}$, and $E(\text{SrO})^{\text{GGA}}$ are the total energy values of NM SrFeO$_3$, bcc FM Fe, and SrO respectively, we acquire $p(\text{O}_2)^{\text{GGA,min}} = 8.08 \times 10^{-14}$ atm at $T = 700$ K. As to the GGA + U situation, the 0.68 eV/O correction is introduced, and the lower limit $p(\text{O}_2)^{\text{GGA}+U,\text{min}}$ is calculated to be 9.86
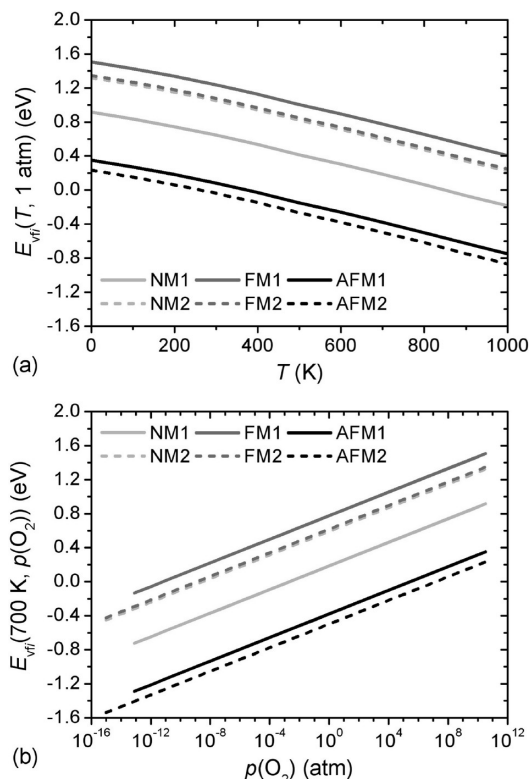
Figure 10. O-vacancy formation energy $E_{vfi}(T, p(O_2))$ ($i = 1, 2$) from SrFeO$_3$ to SrFeO$_{2.875}$ with different magnetic cases (NMi, FMi, and AFMi) at (a) $p(O_2) = 1$ atm and (b) $T = 700$ K.

$\times$ 10^{-16} atm at $T = 700$ K, according to 1/2 $k_B T$ ln($p(O_2)^{GGA+U,\min}/p^o(O_2)$) = 1/2 $E(SrFeO_3)^{GGA+U}$ − 1/2 $E(Fe)^{GGA+U}$ − 1/2 $E(SrO)^{GGA}$ − 1/2 $E(O_2)^{GGA}$ − 1/2 $\Delta G(\Delta T, p^o(O_2), O_2)$ − 0.68 eV, where $E(SrFeO_3)^{GGA+U}$ and $E(Fe)^{GGA+U}$ are the total energy values of NM SrFeO$_3$ and bcc FM Fe SrO, respectively, calculated at the GGA + U level. It is evident that the $E_{vfi}(T, p(O_2))$ decreases with the increasing T or decreasing $p(O_2)$. At $p(O_2) = 1$ atm for the NM structure, the $E_{vf1}(T, p(O_2))$ reaches zero around $T = 850$ K, and $E_{vf2}(T, p(O_2)) = 0.22$ eV at 1000 K. Regarding the NM cases at $T = 700$ K, the oxygen-vacancy formation becomes thermodynamically favorable (exothermic) at $p(O_2)$ below 2 $\times$ 10^{-3} atm within the GGA framework (eq 3) or 3 $\times$ 10^{-9} atm at the GGA + U level with the 0.68 eV/O correction (eq 4). Compared with the data of other systems,[18,29] such ease of oxygen-vacancy formation in SrFeO$_3$ accounts for its potentially wide application to catalysis or electrocatalysis (e.g., solid oxide fuel cells), oxygen separation, and sensors.[30]

In view of the difficulty in evaluating magnetic systems using US-PPs,[31] we have chosen additionally another Fe US-PP including 3s and 3p semicore states with 16 e in total to check some results, and general agreement has been observed. The total energy value of FM SrFeO$_3$ obtained with this PP within the GGA scheme is 1.316 and 0.269 eV/fu lower than those in the NM and AFM cases, respectively, and the corresponding values at the GGA + U level are 2.446 and 0.278 eV/fu. The atomic positions of SrFeO$_{2.875}$ relaxed with this PP exhibit only slight divergences from those in Table 1, and the acquired E_{vf1} values are 0.35, 1.59, and 1.00 eV for the AFM, FM, and NM configurations, respectively,

agreeing very well with the preceding ones. The corresponding E_{vf2} values of 0.07, 1.29, and 1.09 eV here display seemingly large deviations from the aforementioned results, but considering the around 0.09 eV uncertainty of DFT methods,[32] such differences might be acceptable.

4. Conclusions

Using the PW-PP DFT method, we have calculated the structural, electronic, and magnetic properties of SrFeO$_{2.875}$ within the GGA ($+ U$) frameworks, taking NM, FM, and AFM patterns into account. By referring to SrFeO$_3$, the vacancy-induced changes have been clarified.

Beyond the NN Fe3, Sr2, and O7 atoms, Sr1, Fe2, O4, and Fe1 in SrFeO$_{2.875}$ have undergone observable positional modifications during the relaxation as well though the population analyses reveals that the oxygen-vacancy formation transfers charges predominantly to the Fe3 atoms, especially the $3d_{z^2}$ and $3d_{x^2 - y^2}$ orbitals for the NM configuration. This also indicates that the constrained relaxation (within the NN region) is insufficient to determine the structures and energetics of this species.

The reduced DOS at E_F for NM SrFeO$_{2.875}$ has been reflected only slightly at the GGA + U level, while the vacancy-caused localization of the states at E_F has been verified by the magnetic moments and LDOS results. The states of SrFeO$_{2.875}$ at E_F are found to be π^* for the NM structure and σ^* for the FM and AFM patterns.

With the correction of the O$_2$ overbinding error using the value in the literature,[5] the NM and FM cases create similar values of the oxygen-vacancy formation energy at the GGA + U level, which are much smaller than those of many other oxides, indicative of the easy reduction of SrFeO$_3$. The values in the AFM cases are extremely small, implicating the extraordinary ease of the transform under such a magnetic pattern.

Supplementary calculations using an alternative Fe US-PP involving 3s and 3p semicore states have yielded generally consistent results, indicating that treating only 3d as the semicore states with NLCC seems adequate to represent the present systems.

References

(1) Tsujimoto, Y.; Tassel, C.; Hayashi, N.; Watanabe, T.; Kageyama, H.; Yoshimura, K.; Takano, M.; Ceretti, M.; Ritter, C.; Paulus, W. *Nature* **2007**, *450*, 1062–1065.

(2) Koslowski, T. *Phys. Chem. Chem. Phys.* **1999**, *1*, 3017–3023.

(3) Goodenough, J. B. *J. Appl. Phys.* **1966**, *37*, 1415–1422.

(4) Shein, I. R.; Kozhevnikov, V. L.; Ivanovskii, A. L. *J. Phys. Chem. Solids* **2006**, *67*, 1436–1439.

(5) Wang, L.; Maxisch, T.; Ceder, G. *Phys. Rev. B* **2006**, *73*, 195107.

(6) Jaya, S. M.; Jagadish, R.; Rao, R. S.; Asokamani, R. *Phys. Rev. B* **1991**, *43*, 13274–13279.

(7) Baroni, S.; Dal Corso, A.; de Gironcoli, S.; Giannozzi, P.; Cavazzoni, C.; Ballabio, G.; Scandolo, S.; Chiarotti, G.; Focher, P.; Pasquarello, A.; Laasonen, K.; Trave, A.; Car, R.; Marzari, N.; Kokalj, A. http://www.quantum-espresso.org (accessed July 10, 2009).

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

(9) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *13*, 5188–5192.

(10) Marzari, N.; Vanderbilt, D.; de Vita, A.; Payne, M. C. *Phys. Rev. Lett.* **1999**, *82*, 3296–3299.

(11) Shanno, D. F.; Phua, K. -H. *Math. Program.* **1978**, *14*, 149–160.

(12) Blochl, P. E.; Jepsen, O.; Anderson, O. K. *Phys. Rev. B* **1994**, *49*, 16223–16233.

(13) Löwdin, P. -O. *J. Chem. Phys.* **1950**, *18*, 365–375.

(14) Bader, R. F. W.; Beddall, P. M. *J. Chem. Phys.* **1972**, *56*, 3320–3329.

(15) Henkelman, G.; Arnaldsson, A.; Jónsson, H. *Comput. Mater. Sci.* **2006**, *36*, 354–360.

(16) Choi, Y.; Mebane, D. S.; Lin, M. C.; Liu, M. *Chem. Mater.* **2007**, *19*, 1690–1699.

(17) Wyckoff, R. W. G. *Crystal Structures*, 2nd ed.; John Wiley & Sons, Inc.: New York, 1963; Vol. 1, pp 10, 16, 90.

(18) Astala, R.; Bristowe, P. D. *Comput. Mater. Sci.* **2001**, *22*, 81–86.

(19) Matar, S. F. *Prog. Solid State Chem.* **2003**, *31*, 239–299.

(20) Kokalj, A. *Comput. Mater. Sci.* **2003**, *28*, 155–168. Code available from http://www.xcrysden.org/ (accessed July 10, 2009).

(21) Anderson, P. W. *Phys. Rev.* **1958**, *109*, 1492–1505.

(22) Shein, I. R.; Shein, K. I.; Kozhevnikov, V. L.; Ivanovskii, A. L. *Phys. Solid State* **2005**, *47*, 2082–2088.

(23) Pople, J. A.; Gordon, M. H.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622–5629.

(24) Hammer, B.; Hansen, L. B.; Nørskov, J. K. *Phys. Rev. B* **1999**, *59*, 7413–7442.

(25) Reuter, K.; Scheffler, M. *Phys. Rev. B* **2001**, *65*, 035406.

(26) Di Valentin, C.; Pacchioni, G.; Selloni, A. *Chem. Mater.* **2005**, *17*, 6656–6665.

(27) Kuwabara, A.; Tanaka, I. *J. Phys. Chem. B* **2004**, *108*, 9168–9172.

(28) Hodges, J. P.; Short, S.; Jorgensen, J. D.; Xiong, X.; Dabrowski, B.; Mini, S. M.; Kimball, C. W. *J. Solid State Chem.* **2000**, *151*, 190–209.

(29) Fabris, S.; Vicario, G.; Balducci, G.; de Gironcoli, S.; Baroni, S. *J. Phys. Chem. B* **2005**, *109*, 22860–22867.

(30) Peña, M. A.; Fierro, J. L. G. *Chem. Rev.* **2001**, *101*, 1981–2017.

(31) Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*, 1758–1775.

(32) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974–12980.

JCTC Journal of Chemical Theory and Computation

Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer

Roland Schulz,[†,‡,¶] Benjamin Lindner,[†,‡,¶] Loukas Petridis,[†,¶] and
Jeremy C. Smith*[,†,‡,¶]

*Center for Molecular Biophysics, Oak Ridge National Laboratory, 1 Bethel Valley
Road, Oak Ridge, Tennessee 37831, Department of Biochemistry and Cellular and
Molecular Biology, University of Tennessee, M407 Walters Life Sciences 1414
Cumberland Avenue, Knoxville, Tennessee 37996, and BioEnergy Science Center,
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831*

Abstract: A strategy is described for a fast all-atom molecular dynamics simulation of multimillion-atom biological systems on massively parallel supercomputers. The strategy is developed using benchmark systems of particular interest to bioenergy research, comprising models of cellulose and lignocellulosic biomass in an aqueous solution. The approach involves using the reaction field (RF) method for the computation of long-range electrostatic interactions, which permits efficient scaling on many thousands of cores. Although the range of applicability of the RF method for biomolecular systems remains to be demonstrated, for the benchmark systems the use of the RF produces molecular dipole moments, Kirkwood G factors, other structural properties, and mean-square fluctuations in excellent agreement with those obtained with the commonly used Particle Mesh Ewald method. With RF, three million- and five million-atom biological systems scale well up to ∼30k cores, producing ∼30 ns/day. Atomistic simulations of very large systems for time scales approaching the microsecond would, therefore, appear now to be within reach.

1. Introduction

Molecular dynamics (MD) simulation is a powerful tool for the computational investigation of biological systems.[1] Since the first MD study of a protein in 1977, which simulated <1000 atoms for <10ps,[2] significant progress has been made in the time and length scales accessible, and it is now common to probe systems of ∼10^5 atoms on time scales of ∼100 ns. This increase in scope has allowed many processes of biological interest to be characterized. However, there is clear interest in further extending both the time and the length scales beyond those currently accessible.

Recent algorithmic[3−6] and hardware developments have allowed MD studies to be extended to multimillion-atom systems.[7−9] Current supercomputers, such as the "Jaguar"

Cray XT5 at Oak Ridge National Laboratory used for the present study, are beginning to assemble over 1×10^5 cores and in this way reach petaflop nominal speeds. However, the challenge for MD, and other applications, is to achieve efficient scaling up to ∼1×10^4 to 1×10^5 cores, i.e., the simulations are limited by the parallel efficiency of the MD algorithms.

The computationally most demanding part of MD simulation of biological systems is the treatment of long-range interactions, which in nonpolarizable force fields is represented by Coulomb and van der Waals (Lennard-Jones) terms.[10] While the van der Waals interaction is almost always truncated at a cutoff distance R_{vdw}, the electrostatic interaction extends to longer ranges. A common method to treat the electrostatics is to directly calculate the Coulomb interaction for any pair of atoms separated by less than another cutoff distance R_{coul} and, outside this distance, to calculate the interactions with the Particle Mesh Ewald[11,12] (PME) method

* Corresponding author. E-mail: smithjc@ornl.gov.
† Oak Ridge National Laboratory.
‡ University of Tennessee.
¶ BioEnergy Science Center.

Biological Molecular Dynamics Simulation

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2799**

(assuming periodic boundary conditions are applied to the system). By using an Ewald summation to split the Coulomb interaction into a short-range part that converges quickly in real space and a long-range part that converges quickly in reciprocal space, the PME method reduces the computational cost of N particles interacting with each other from $O(N^2)$ to $O(N \ln N)$. The reciprocal space sum is performed by using the fast Fourier transformation (FFT).

Most MD calculations have been performed using PME. Although PME suffers from artifacts introduced by the periodicity,[13−17] it is fast on a small number of processors, and FFT-based electrostatics methods are also very successful on special purpose hardware. However, on massively parallel computers, the full electrostatic treatment via the PME method presents a performance barrier, arising from the state-of-the-art implementation of PME, which requires two FFT steps. The FFT algorithm in turn requires one or two global transposes which, on a message passing system, is inherently limited by the bandwidth and the latency of the network. As more nodes are used to simulate a system, each MD time-step can be calculated faster, and thus, the time between communications becomes shorter. If the time for the global transpose is of the same order of magnitude as the computation time, then the required communication becomes a bottleneck for the parallel efficiency. The same reasoning explains why, when running on the same number of cores, the parallel efficiency of a large system (e.g., 1×10^6 atoms) is much better than that of a small system (e.g., 1 000 atoms): the time needed to compute a single time step on a single processor is much longer in case of a large system. Therefore, for a large system, many more cores can be used before the communication bottleneck occurs. As a result, larger systems can often be simulated at about the same speed (in ns/day) as smaller systems.

An alternative method to PME, that avoids the electrostatics bottleneck, is the reaction field (RF).[18−21] In RF it is assumed that any given atom is surrounded by a sphere of radius, r_c, again within which the electrostatic interactions are calculated explicitly. Outside the sphere, the system is treated as a dielectric continuum. The occurrence of any net dipole within the sphere induces a polarization in the dielectric continuum, which in turn interacts with the atoms inside the sphere. Due to the approximate treatment of electrostatics outside the cutoff sphere, the RF method in principle does not yield accurate results for systems that contain full charges. Nevertheless, the RF model allows the replacement of the infinite Coulomb sum by a finite sum modified by the reaction field and, therefore, limits the parallel scaling less than the PME method.

Many simulations have used the RF model for electrostatics in the past.[14,22−30] Testing of the accuracy of RF for charged biomolecular systems has also been performed. A study on the thermodynamics of the folding of a short peptide with a net charge of −1 found that a RF simulation produces results similar to those with the PME calculations and experiments:[27] the free energy surfaces derived from both the RF and the PME calculations exhibit the same single minimum, which corresponds to a β-hairpin, the experimentally determined native state of the peptide. The main

difference between the PME and the RF results concerned the structures of the less frequently sampled unfolded configurations. The validity of the RF has also been discussed in an extensive examination of the effects of force fields and electrostatics treatments on the secondary structure propensity and the sampling in the peptide folding simulations.[30] It was found that the electrostatics treatment has little effect on the folding characteristics of the peptides, with the PME exhibiting a slightly lower rmsd relative to the native state, but the RF had slightly better sampling. Earlier studies on a highly charged protein crystal,[23] a RNA hairpin in solution,[24] and a small, highly charged globular protein[25] also found that RF produced similar structures[23−25] and conformational spaces sampled[24,25] compared to PME. However, the studies in refs 23−25 involve relatively short trajectories (between 1.5 to 5 ns) and, therefore, do not provide a rigorous test of the use of the RF in representing longer time scale dynamics. In ref 30, it was shown that large conformational changes in proteins in solution can occur on the microsecond time scale, and these might, in principle, be sensitive to electrostatics.

The present paper outlines a strategy for fast and accurate all-atom simulation of multimillion-atom biomolecular systems that do not contain charged groups. The benchmark systems used in the present study are cellulose in water and models of lignocellulosic "biomass". Lignocellulosic biomass is a complex material composed of crystalline cellulose microfibrils laminated with hemicellulose, pectin, and lignin polymers.[31] In recent years, there has been a revived interest in biomass structure, as biomass offers a potentially abundant and cheap source of sugar for industrial biofuel production.[32] Due to its complexity, lignocellulose poses significant challenges to MD simulation. Among these are the characteristic length scales (Å−μm) and time scales (ns−μs and beyond) of events pertinent to the recalcitrance of biomass to hydrolysis into sugars.[32] To access these length and time scales, standard MD protocols must be modified to scale up to massively parallel machines. Two technical problems are addressed. First, we compare the accuracy of MD using PME and RF on the benchmark systems, and, second we examine the scaling of MD of large systems on a petascale supercomputer.

The present comparative studies show that the examined properties derived using PME are well reproduced using the computationally less demanding method of RF. Scaling benchmarks on multimillion-atom systems show that the use of the RF drastically improves the parallel efficiency of the algorithm relative to PME, yielding ∼30 ns/day. Consequently, microsecond time scale MD of multimillion-atom biomolecular systems appear now within reach.

2. Methods

2.1. Simulation Setup. The simulations were performed using cellulose[33] and lignin[34] force fields parametrized for the CHARMM energy function[35] using GROMACS 4.0.4[4] as the MD software. The validation of the use of the RF is

Table 1. Sets of Benchmark Simulations Performed[a]

simulation index	electrostatic treatment
1	PME with switch
2	PME with cutoff
3	RF
4	shift

[a] Each set comprises five 20 ns trajectories initiated from the same structure but with different initial velocity distribution.
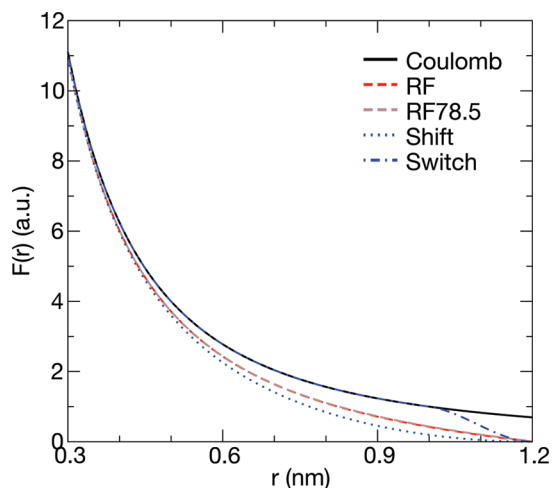

Figure 1. Coulomb force as a function of the distance between opposite charges. Coulomb is the Coulombic force without modification. RF is the reaction field with $\varepsilon = \infty$ outside the cutoff radius. For RF78.5, $\varepsilon = 78.5$ outside the cutoff radius. Shift and switch are computed as described in the user manual of GROMACS.[36] The switch distance after which the Coulomb function is altered is 1nm.

particularly important in the present benchmark systems, since the force fields were not parametrized using the RF method.

The GROMACS simulations were performed with the electrostatic treatments RF, PME-cutoff, PME-switch, and shift and switch (see Table 1). The analytical expression for electrostatic potential, V_{rf} with the RF method is

$$V_{rf} = \left(1 + \frac{(\varepsilon - 1)r^3}{(2\varepsilon + 1)r_c^3}\right)r^{-1} - 3\frac{\varepsilon}{r_c(2\varepsilon + 1)} \quad (1)$$

where ε is the dielectric constant outside the radius r_c, and r is the distance separating two charges. In the present work, we employ $\varepsilon = \infty$. This has the advantage that the force is continuous at the cutoff distance, whereas RF with a finite dielectric constant is subject to errors due to force discontinuity. Apart from this significant improvement, the Coulomb forces of the RF with $\varepsilon = \infty$ and 78.5 are almost identical (see Figure 1).

In switch and shift a function S is added to the Coulomb force F_c, giving a total force $F_t = F_c + S$. S is a third-degree polynomial acting over interatomic distances r where $R_1 < r < R_{coul}$ and is zero otherwise, R_{coul} being the cutoff radius.[36] R_1 is zero for shift and corresponds to the switch-on distance for the switch method (in this study $R_1 = 0.8$ nm and $R_{coul} = 1.2$ nm, more details on the simulation parameters follow in the next paragraphs). The polynomial is constructed so that $S(r_1) = S'(r_1) = F_t(r_c) = F_t'(r_c) = 0$. The switch, shift,

RF, and Coulomb functions are shown in Figure 1. The switch electrostatics was immediately found to produce severe artifacts, including a strong suppression of the fluctuations of the heavy atoms of the cellulose. Consequently the switch simulation was not further considered for detailed analysis. We suspect the switch-induced errors to have been enhanced by the periodicity in the fibril.

In the so-called "atom-based" cutoff, the electrostatic interactions between atoms that are separated by less than the cutoff distance, are explicitly calculated. Some MD codes, such as CHARMM[37] and GROMACS[4] but not NAMD,[3] subdivide molecules into neutral "charge groups". These charge groups are composed of a small number of covalently bonded atoms, such as a methyl group or the HO−C−H moiety of cellulose (see Figure 7). A group-based cutoff can then be defined in which the electrostatic interactions between all atoms belonging to two "charge groups" are calculated explicitly, as long as the geometric centers of the two groups are separated by less than the cutoff distance.

The introduction of the cutoff distance, r_c, is accompanied by a meaningful reduction in the computational cost of the electrostatics only if the list of atoms that are separated by less than the cutoff distance is not calculated at every step. A list of neighbor atoms is created containing all pairs separated by the neighbor list distance, R_{ne}, which is greater than the cutoff distance r_c. The list of neighbors is calculated and updated after n_{list} steps, with n_{list} usually taking values between 10 and 24. The neighbor search distance, R_{ne}, must be large enough to ensure that undesirable events in which atoms jump abruptly in or out of another atom's cutoff sphere do not occur. This would happen, for example, if two atoms are initially separated by $> R_{ne}$, but after the n_{list} simulation steps move significantly closer, to a distance $< r_c$.

The crystal structure of the I_β allomorph of cellulose[38] was used. This cellulose structure has two chains per triclinic unit cell, which will be referred to as the "origin" and the "center" chains. The shape of the fiber was hexagonal.[39] Figure 2 shows structural details of the model. Details on the models of the lignin molecules are presented elsewhere.[40]

For the simulations in which the effects of varying the electrostatic model were examined, the cellulose contained 80 monomers per chain (36 chains) and was solvated with 70 656 TIP3P[41] water molecules, producing a model totaling 272 556 atoms. A triclinic box was used with a 60° angle between the two short box vectors. The systems were equilibrated for 1ns and then simulated for 20 ns with a time step of 2 fs. For each simulation setup, five simulations with different initial velocities were run. Neighbor searching was performed every 10 steps. All bonds were constrained using LINCS[42] (order: 3, iterations: 2). Temperature coupling was performed with the Nosé−Hoover[43] algorithm ($\tau = 1$ps) during equilibration and the Berendsen[44] algorithm ($\tau = 0.1$ps) during production. Pressure coupling was performed with the Berendsen algorithm (semi-isotropic, $\tau = 4$ ps) during equilibration and the Parrinello−Rahman[45] (isotropic, $\tau = 4$ ps) algorithm during production.

For all lignocellulose simulations with the RF, a dielectric constant of $\varepsilon = \infty$ and a group-based cutoff were employed with the so-called reaction field-zero method, which uses
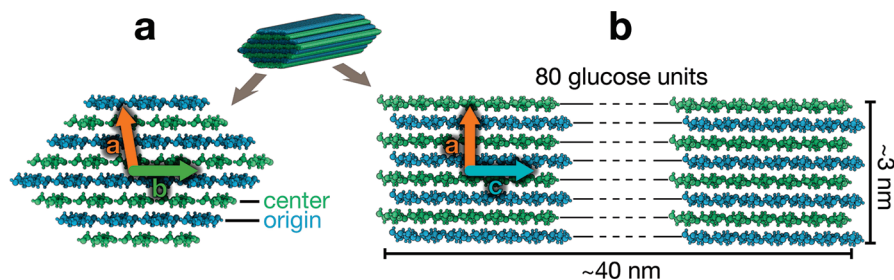
Figure 2. The model of the simulated cellulose fibril showing: (a) the cross-section and (b) a side perspective. The fibril consists of 18 origin chains (blue) and 18 center chains (green). The axes of the unit cell are also indicated.

spline interpolated tables instead of the analytical expression.[36] For the RF, shift, and PME with switch runs a neighbor list-search distance of 1.5 nm, a electrostatic and VDW cutoff of 1.2 nm, and a switch distance of 0.8 nm were used. For PME without switch a neighbor list-search distance of 1.2 nm, a electrostatic cutoff of 1.2 nm, a VDW cutoff of 1.0 nm, and a VDW switch distance of 0.8 nm were used.

In a first analysis step, the simulations were inspected visually. This inspection showed that a strong artifact can arise in the case where only a small buffer region is employed between the cutoff radius and the neighbor list-search distance. To determine the optimal width of the buffer region, a series of simulations was performed varying the width from 0 to 0.3 nm in 0.1nm steps. Simulations using non-PME electrostatics and buffer regionsof <0.3 nm exhibited artificial deterministic linear translation of whole cellulose fibers along their axes with a speed of ~1m/s. Thus, for all further analysis, a buffer region of 0.3 nm was used for non-PME electrostatics.

For the supercomputing performance comparisons, a system was constructed of lignocellulosic biomass containing 52 lignin molecules each with 61 monomers, the same cellulose fibril as described above and 1 037 585 TIP3P water molecules, totaling 3 316 463 (or 3.3 million) atoms. All simulation settings, apart from bond constraints, were the same as the RF settings for cellulose given above. All bonds involving hydrogens were constrained with LINCS[46] (order: 4, iterations: 1). For the sole purpose of extending the scaling tests to a larger system, an additional model system containing 64 000 dipeptide molecules (GLY-PRO) and 1 280 000 water molecules, totaling 5 376 000 atoms, was also constructed. The system was simulated with the same protocol and parameters as the 3.3 million atom lignocellulose system. The detailed system setup is described elsewhere.[47] For the PME simulations in Figure 9, the NAMD multiple time step method was used, in which the long-range electrostatics is computed only every third step and, in addition, a smaller buffer was used.

Topologies were generated in CHARMM[48] and converted using a locally modified version of psfgen[49] (see Supporting Information for details). The correctness of the converted topology and force field was checked by comparison with CHARMM and NAMD (see Supporting Information). All analysis was performed using the tools provided by GROMACS.[50,51] The NAMD trajectories were converted for analysis to GROMACS format and reordered with catdcd.[52] Molecular drawings were made with QuteMol.[53]

2.2. Supercomputer Performance Measurements. The performance tests were run on JaguarPF, a Cray XT5 massively parallel processing (MPP) computer with over 150 000 Opteron 2.3 GHz cores. JaguarPF has a LINPACK performance of over one petaflop and a SeaStar 2+ interconnect. The internal timings of GROMACS 4.0.4 and NAMD (CVS version) were used. Both CHARM++ and NAMD were built with the Cray-XT4 optimized settings available in the source code. For all tests, all eight cores of each node were used. No special benchmark partition was setup, and as a result, the benchmarks were subject to the regular placement of nodes by the scheduler. The Cray-XT scheduler does not allow consecutive blocks of nodes to be allocated, and the run time varied depending on the node placement. The IO time was included in the benchmarks.

For RF with GROMACS, all parameters were as described in the system setup Section 2.1. For PME with NAMD, a neighbor list-search distance of 1.35 nm, a multiple time step method with a full electrostatic frequency of 3 and a steps-per-cycle of 24, and a grid spacing of 0.13 nm was used (this relatively large spacing was used to ensure good performance by the PME/NAMD simulation). A variation was observed in the speed (in ns/day) of the benchmark runs that used the same number of cores. This variation was caused by the reading input/writing output (IO time) and the task placement.

The IO time was found to be impacted by latency problems caused by Lustre scaling (due possibly to the meta data server). The currently available profiling data do not conclusively identify the relative contribution of Lustre and node placement to the variation of the performance. We have chosen to consider only the best three times (although all 12 times are given in the Supporting Information), as the best three times reflect the optimal performance of the Cray XT5, i.e., when the Lustre system and node placement do not impede the performance of the MD codes. Each MD run was limited to a wall clock time of 10 min. For the thermostat and barostat, the total energy and virial were computed every 10 steps in RF/GROMACS and every 24 steps in PME/NAMD. The calculation of the total energy/virial requires a MPI_Allreduce communication, and therefore, more frequent updates would limit the scaling. For the domain decomposition (DD), the 12 288 cores were arranged in a 3D $96 \times 16 \times 8$ grid. The load balancing works by changing the volume of the DD cells relative to each other. For the minimum DD cell length, 0.77, 0.68, and 0.34 of the average length were used for X, Y, and Z respectively.

3. Results

3.1. Comparison of Simulations with Different Electrostatic Methods. As will be discussed in Section 3.2, fast MD simulation of the 3.3 million atom lignocellulose system can be obtained using the RF method for treating the electrostatic interactions. This section is devoted to examining the accuracy of RF for the biomass test systems. For this, structural and dynamical properties are compared in simulations using different electrostatic methods for both the solvent and the solute. The particular choice of properties for comparison in the case of the solute is based on their structural importance and anticipated sensitivity to possible electrostatic artifacts.

In order to investigate the dependence of dynamical properties on the chosen electrostatics method, the set of MD simulations listed in Table 1 was analyzed. The system of a cellulose fibril in aqueous solution, i.e. without lignin, was chosen for this comparison. Lignin was omitted since significantly longer trajectories would be required for the convergence of dynamical properties due to its amorphous character, thus, complicating the comparison.

Quantities were calculated that were expected to be particularly sensitive to electrostatics. Two functions probing the electrostatic-induced structure and dynamics are the total dipole moment of the fiber and the Kirkwood function between the dipoles of different chains, the latter providing information on the distance-dependent correlation of molecular dipoles. Finally, three specific dihedral angles were selected for comparison due to their structural importance in cellulose.

3.1.1. Solvent. The applicability of the reaction field method for pure water has been discussed in literature (see for example refs 22, 26, and 28). Different implementation schemes exist which can be classified into two groups, namely atom and charge-group based truncation methods.

A suitable function for probing the structural artifacts of water is the Kirkwood factor, $G_k(r)$, which is given by[54]

$$G_k(r) = \sum_{r_{ij} < r} \frac{\vec{\mu}_i \vec{\mu}_j}{|\vec{u}|^2} \qquad (2)$$

where $\vec{\mu}_i$ and $\vec{\mu}_j$ are the electric dipole moments of water molecules i and j, respectively. $G_k(r)$ is a measure for the orientational ordering of the dipole moments of the solvent molecules. In earlier work, significant discrepancies were found between PME and RF78.5 (i.e., RF with $\varepsilon_r = 78.5$) in the simulation of bulk water.[26,28] We, therefore, investigated the cause of these discrepancies. To reproduce the earlier results, a simulation setup of bulk water as described in ref 28 was constructed. The resulting Kirkwood factor for water is shown in Figure 3 for four distinct electrostatic treatments with this setup: (a) PME, (b) RF with an infinite dielectric constant ε and group-based cutoff, termed RF in Figure 3, (c) RF with $\varepsilon = 78.5$ and group-based cutoff and (d) RF with $\varepsilon = 78.5$ and an atom-based cutoff. For RF (b) interpolation of tabulated values was used and for RF78.5 (c and d) the analytical expression of RF was used directly.

The RF method with $\varepsilon = \infty$ shows the best overall agreement with PME in terms the of residual difference. For
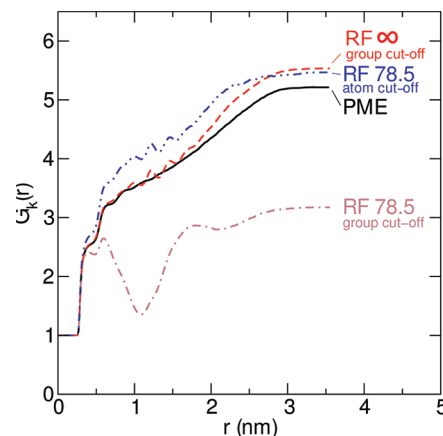
Figure 3. Distance-dependent Kirkwood factor (eq 2) for (a) PME, (b) RF with $\varepsilon = \infty$ and group-based cutoff, (c) RF with $\varepsilon = 78.5$ and group-based cutoff, and (d) $\varepsilon = 78.5$ and atom-based cutoff (where ε is the dielectric constant outside the cutoff radius). RF 78.5 with group-based cutoff produces artifacts that are corrected by using an atom based cutoff and updating the neighbor list at each step.

group-based cutoff RF78.5, the Kirkwood function is very different from PME with a deep minimum around the cutoff distance, agreeing with the previous observations.[26,28] We performed several additional simulations (not shown) to find the reason for this difference of $G_k(r)$ for RF78.5. It turns out that the discrepancy arises from the combination of the neighbor-list search with the behavior of the analytical RF expression (eq 1). It is possible to simulate with an atom-based neighbor list, by splitting the water molecule into three charge groups, as proposed originally in ref 22. Using this atom-based cutoff and updating the neighbor list at each step, the disagreement between RF78.5 and PME can be reduced, yielding a general agreement on the shape of the curve. However, in what follows, all lignocellulose simulations were performed using an infinite dielectric constant and a group-based cutoff.

GROMACS calculates the electrostatic interaction for all atom pairs included in the neighbor list. The distance between atoms for which the electrostatic interaction is calculated can be larger than the cutoff distance in two cases: for a group-based cutoff or a neighbor-list search with frequencies <1/step. In the former case, only the group center needs to be within the cutoff distance for all atoms of the group to be included in the list. In the latter, it is sufficient for the atom to be within the distance at the time of the neighbor-list search, even if it moves outside afterward. The analytical equation (eq 1) of RF78.5 is negative for distances longer than the cutoff distance. Consequently, we conclude that the earlier observed difference between RF and PME[26,28] arises from the negative interaction of atom pairs separated by distances longer than the cutoff distance, caused by the group-based neighbor-list search. Using a spline interpolated table for RF78.5 instead of the analytical expression, as reaction field-zero does by default, allows the interaction for these distances to be set to zero.

3.1.2. Solute/Cellulose. Cellulose was chosen as a benchmark for the solute since it has a high degree of order and,
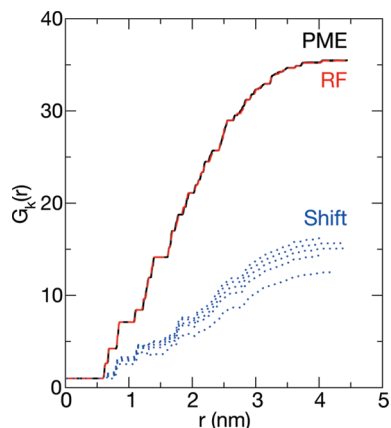
Biological Molecular Dynamics Simulation

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2803**

Figure 4. Distance-dependent Kirkwood factor (see eq 2). The two sets performed with PME are shown as indistinguishable black/solid lines. The RF set is red/dashed lines, and the shift set is blue/dotted lines. The profiles of simulations with PME and RF are almost identical, implying very good agreement between the two methods.

therefore, is anticipated to be adequately sampled over the time scales probed.

3.1.3. Dipole Correlation. In the first comparison, shown in Figure 4, the Kirkwood factor, $G_k(r)$, of cellulose is presented. The form of the Kirkwood factor corresponds to that given by Equation 2,[54] where $\vec{\mu}_i$ and $\vec{\mu}_j$ are now the electric dipole moments of glucose chains i and j, respectively, and r_{ij} then becomes the distance between their centers of mass. In this way, $G_k(r)$ is a measure for the orientational ordering of the dipole moments of the cellulose chains in the fibril. It is clear from Figure 4 that the RF method is in very good agreement with the PME method, contrasting with shift in which $G_k(r)$ is much lower. The spread of the shift profiles arises from differences between the individual simulations in the set.

3.1.4. Total Dipole Moment. A further useful test for global changes in dipolar correlation is the total dipole moment for a given macromolecule. Therefore, this should serve as a further benchmark for the accuracy of the electrostatic methods. As seen in Figure 5, the conclusions drawn from this comparison agree with those from the Kirkwood G factor: RF and PME show similar features (RF yielding a total dipole moment about 1% lower than PME), whereas the shift method results in a 50% lower magnitude and has slower convergence.

3.1.5. RMSF and RMSD. General dynamical properties examined include the root-mean-squared fluctuations (RMSF) and the modes resulting from principal component analysis (PCA). The RMSF is a measure of the fluctuations of atoms around their equilibrium structure, and PCA provides information on the major collective modes of motion. Both properties are commonly calculated in biomolecular simulations and were, therefore, checked for reproducibility.

Figure 6 shows the difference between the time-averaged RMSF of each atom in the cellulose fibril computed with the RF method minus the RMSF computed with the PME method ($\Delta(\text{RMSF})_{\text{RF}}$). Shown also is the RMSF difference between the shift and PME methods ($\Delta(\text{RMSF})_{\text{shift}}$). The overall good agreement between the RF and the PME is observed once more: the RF enhances fluctuations slightly (with a more pronounced effect for the atomic indices in the range 35 000–40 000), but
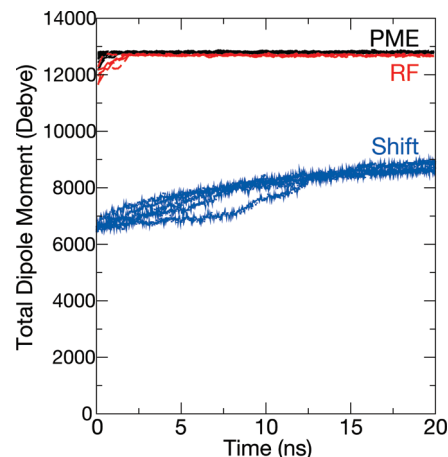
Figure 5. Total dipole moment of the cellulose fibril plotted. The two sets performed with PME are indistinguishable black/solid lines, the RF set in red/dashed lines, and the shift set in blue/dotted lines. The profiles of simulations with PME and RF are almost identical.

the shift leads to a much stronger deviation from the PME behavior. In contrast, the rmsd as a function of time, shown in the Supporting Information (Figure S1), shows little difference between the electrostatic treatments. Also in the Supporting Information it is shown that the amplitudes of the eigenvalues obtained from the PCA of the trajectories using the three methods are similar.

3.1.6. Dihedral Angles. The final test focuses on important local structural features of crystalline cellulose. Two sets of dihedrals are examined, as indicated in Figure 7. The particular relation of these dihedrals with respect to cellulose structure is discussed in detail in MD studies of cellulose.[55,56] The ω dihedral (O6−C6−C5−C4) determines the configuration of the primary alcohol group,[55] which affects the hydrogen bonding between adjacent glucose chains within a cellulose fiber and, therefore, is a main determinant for the crystalline phase. When the alcohol lies on the plane of the five-membered glucose ring ($\omega = -60°$ or $\omega = 180°$), single monomers preferentially hydrogen bond to partners within the (010) crystal plane, whereas when the primary alcohol points perpendicular to the five-membered ring plane ($\omega = 60°$) intersheet hydrogen bonds are formed. The Φ and Ψ angles (O5−C1−O1−C4*/C1−O1−C4*−O5*, where * marks atoms on the succeeding monomer) describe the twisting between two consecutive monomers and probe for the frustration in twisting behavior of isolated glucose chains induced by the fiber structure. Unlike the previous properties, these dihedral measures were not necessarily expected to be especially sensitive to differences in electrostatic treatment. They do, however, play an important role in the structure of cellulose. It is, therefore, of interest to determine weather their PMF are not significantly affected by variation of the electrostatic treatment.

The PMFs were calculated according to the equation:

$$W(\theta) = -k_B T \log P(\theta), \qquad \theta = \{\omega, \Phi, \Psi\} \quad (3)$$

where θ is the dihedral angle in question, and $P(\theta)$ is the associated probability distribution. Since the I_β crystal phase of cellulose has two distinct chains per unit cell, a total of six PMF calculations was performed: for each of the three
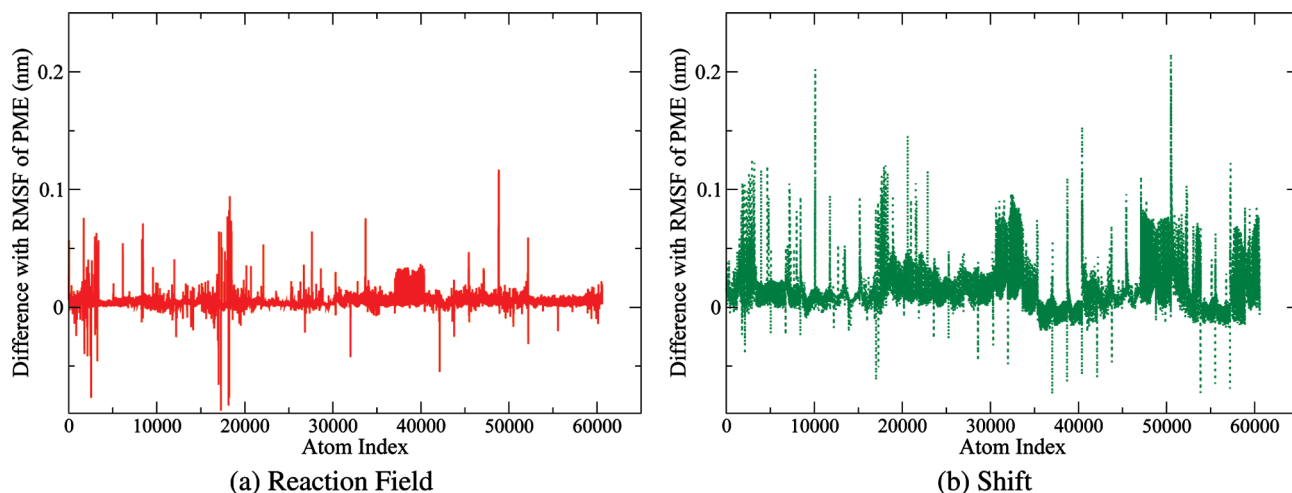
Figure 6. $\Delta(RMSF)_{RF}$ and $\Delta(RMSF)_{shift}$ as defined in Section 3.1.2 for all atoms in the cellulose fibril (atomic index on *x*-axis).
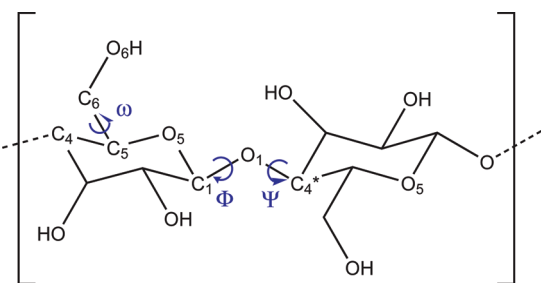

Figure 7. Sketch of cellobiose, the repeating unit of cellulose, indicating the three important dihedrals: the primary alcohol ω dihedral and the Ψ and Φ dihedrals.

dihedrals (ω, Φ, and Ψ), the PMF was calculated for the center and the origin chains. The resulting plots are shown in Figure 8.

The PMFs for the primary alcohol dihedrals follow the same trend as the previous benchmarks, i.e., there is good agreement between the RF and the PME methods but not with the shift method. We note that comparison of the profiles is only meaningful at the relatively low-energy regions that are adequately sampled. In the PMF for the origin chains in 8a, the RF and the PME profiles are almost indistinguishable. However, with the shift method the global minimum moves from 70° to 50°. The difference between the shift and the PME is even more pronounced in the PMF for the center chains (8b), for which the shift introduces a new minimum at −80°, which is only a weak shoulder in the PME calculations.

It is of interest that in the crystal structure of cellulose all primary alcohols have ω = −60°.[38] The transition from ω = −60° to ω = 180° observed during the MD simulation is as expected and has been reported in previous MD studies.[55] The origin of the transition is that the force field employed[33] was parametrized for glucose in water and favors the ω = 180° conformation. Curiously, the shift method appears to "correct" this shortcoming of the force field, and the ω = −60° conformation is populated in the center chains. However, this effect is probably a cancellation of errors. The present test concerns not the accuracy of the force field with respect to experiment, but rather a comparison between the methods for treating long-range electrostatics.

The PMF for the Φ and Ψ dihedrals shows little variation between the simulations using the three electrostatic treatments and is given in the Supporting Information, Figures S3 and S4.

In summary, the RF electrostatics method is found to accurately reproduce simulation results performed with the widely used PME method. From the set of tests performed, it can be, therefore, inferred that no obvious artifacts are introduced by the proposed RF methodology when applied to the simulation of systems with no net charges.

3.2. Scaling. The parallel efficiency of the RF MD simulation is now evaluated by considering strong scaling. In strong scaling, the system size (i.e., here, the number of atoms in the system) is held constant, while the number of cores used varies. The strong scaling of the 3.3 million atom MD simulation of lignocellulose, using the RF on the ORNL "Jaguar" Cray XT5, is shown in Figure 9. For this system, GROMACS scales well to the 12 288 cores and achieves 27.5 ns/day, running at 16.9TFlops. This performance is made possible by the good scaling of the RF, and a fast particle−particle streaming SIMD (single instruction, multiple data) extensions (SSE) compute kernel running for the lignocellulose system at 4GFlops per opteron core. The RF also improves the parallel efficiency of the MD simulation of even larger systems. Figure 10 shows the strong scaling of a 5.4 million atom peptide solution test system. The same production of 28 ns/day is obtained, this time scaling well to 30k cores.

The load balancing is most critical for the scaling to several thousand cores, as the load per volume of each domain decomposition cell is not equal. The primary cause of load imbalance is the difference in computational speed between the solvent and the solute. The higher speed for the solvent arises from a specially optimized water compute kernel and fewer van der Waals interactions for water. We implemented a new way of performing the dynamical load balancing in GROMACS, detailed in the Supporting Information, improving the average load imbalance from 200 to 75% and leading to an overall 44% improvement of the performance. This improvement resulted in the code obtaining the same production (in ns/day) using half the cores that were used
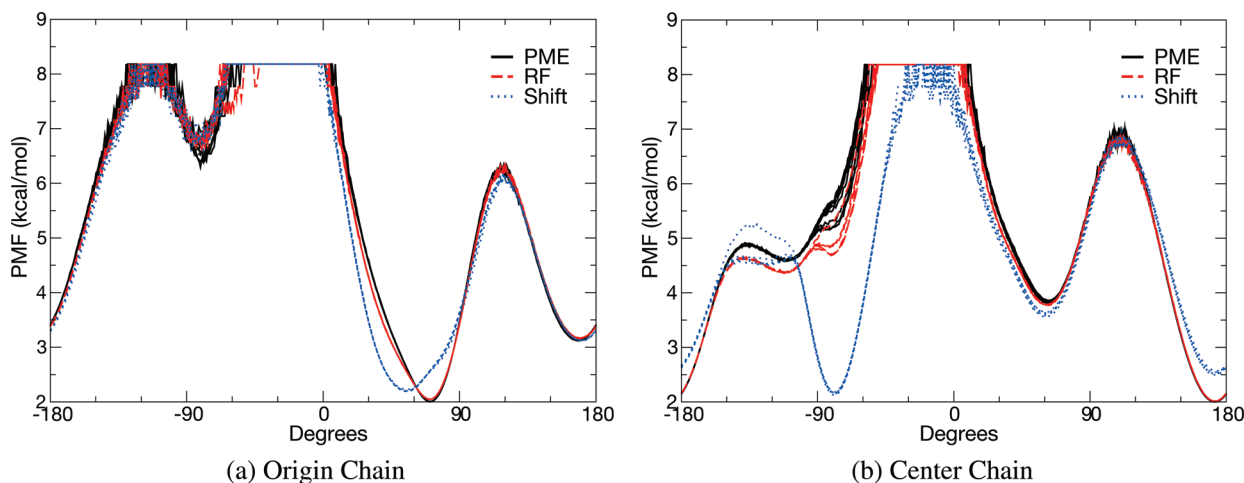
(a) Origin Chain   (b) Center Chain

Figure 8. Potentials of mean force for the primary alcohol dihedral $\omega = O6-C6-C5-C4$: (a) results from all 36 origin chains and (b) results from all 36 center chains.
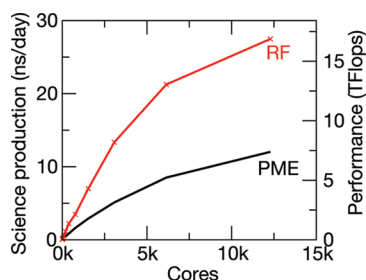

Figure 9. Strong scaling of 3.3 million atom biomass system on Jaguar Cray XT5 with RF. With 12 288 cores the simulation produces 27.5 ns/day and runs at 16.9TFlops. As a comparison, the performance of PME is shown.
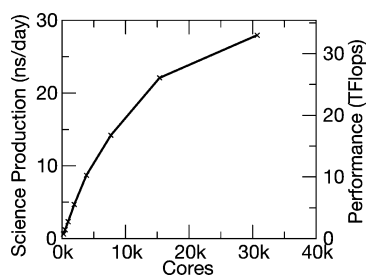

Figure 10. Strong scaling of 5.4 million atom system on Jaguar Cray XT5. With 30 720 cores, 28 ns/day and 33TFlops are achieved.

prior to the improvement. The new implementation of the load balancing is now part of the GIT version of the GROMACS and will also be included in GROMACS 4.1.

To highlight the computational benefit of using the RF, the scaling of a simulation of the 3.3 million lignocellulose system using the PME method is also shown in Figure 9. The PME simulation was run using NAMD, since this MD application is known to have good parallel efficiency.[57] To ensure a "fair" comparison between the two electrostatics methods, some of the parameters of the PME simulation were adjusted to improve its performance (standard 2 fs time step for RF and 6 fs full electrostatics time step and neighbor-list distance update for PME, see Section 2.1 for details). In particular, the reason the RF calculation is faster than the PME at low levels of parallelization is that, on a single processor, the time per step for GROMACS with RF is
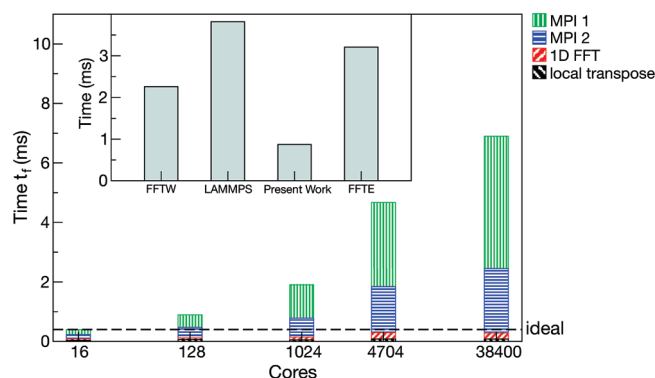
Figure 11. Weak scaling of complex-to-complex FFT on Cray XT5 with FFT implemented as described in Supporting Information, A.3. The 3.3 million atom system requires the $588 \times 128 \times 128$ FFT. The time required to compute one FFT step is represented by t_f.

shorter than for NAMD with PME. However, we stress that the aim of this benchmark is a comparison between the electrostatic treatments and not between the different MD applications. Two different applications were used simply because a direct comparison of simulations using different electrostatics methods with one application is presently not possible: NAMD, which is presently the most scalable code using PME on Cray XT, does not have RF implemented, and GROMACS does not yet have an efficiently scaling PME implemented, with the consequence that PME calculations using GROMACS currently scale up to less than 1 000 cores for large systems (for more details see the Supporting Information).

The significant difference in the parallel efficiency of the PME and the RF electrostatics methods, demonstrated in Figure 9, can be understood by examining the weak scaling of the parallel FFT required for PME, shown in Figure 11. In weak scaling, the ratio of the problem size to the number of cores used in the simulation is held constant. The FFT is a new and improved implementation, the technical details of which are presented in Supporting Information, A.3. The Inset of Figure 11 shows that the new FFT is faster than the FFTs from LAMMPS-FFT,[58] FFTE 4.0,[59] and FFTW 3.2.[60] In ideal weak scaling, the time, t_f, required to perform one

FFT step, indicated by the height of the bars in Figure 11, would remain constant as the number of cores used in the calculation increases from 16 to 38 400. In practice, however, Figure 11 shows that parallel FFT calculations show poor weak scaling, with t_f increasing dramatically on a large number of cores. This increase is a result of the large increase of the required communication time (MPI-1 + MPI-2 in Figure 11). Since, in weak scaling, the number of cores is proportional to the size of the simulated system, Figure 11 demonstrates that the PME method becomes computationally inefficient for large systems.

4. Discussion

This paper presents a strategy for efficient atomistic MD scaling of biological systems on massively parallel super-computers. The key element of the strategy is to compute the long-range electrostatic interactions with the reaction field (RF) method.

In recent years many simulations have been performed using the PME method. This popularity can be attributed to its high accuracy and fast performance. The most efficient way to perform PME calculations is to balance the work so that one-third to one-fourth of the total CPU-time is spent on the PME part.[4] Hence, removing the PME part improves the overall performance of MD calculations by only ~25−33%. However, since RF requires a longer neighbor list, as explained in detail previously, the amount of work required for the direct part increases. Since the number of interactions within the neighbor-list distance increases with the volume and, thus, the third power of the radius, the work increases by 2.0 times (or 3.4 times) when increasing the distance from 1.0 to 1.2 nm (or 1.5 nm). This assumes all interactions in the neighbor-list distance are calculated, as is currently implemented in GROMACS, because selecting interactions is expensive on modern CPU architectures. Thus, when comparing PME with a shorter neighbor list to RF with a longer neighbor list (to ensure accurate results), PME is faster on a small number of processors. This picture, however, changes dramatically when MD simulation is run on a large number of processors, where the PME method displays inherent scaling problems.

The scaling of MD codes is restricted by global communications, i.e., instances when all computer nodes exchange information. Although an improvement in the FFT part of PME speed is reported in the Supporting Information, MD simulations using PME still face weak-scaling problems. While for small systems, containing less than 100k atoms, simulations achieving over 100 ns/day are currently possible,[4,6,61,62] for larger systems the global communication for the FFT (MPI_Alltoall) takes longer than the time available for one time step on a large number of cores. This problem worsens as the size of the system increases further, as the MPI_Alltoall global communication does not exhibit efficient weak scaling.

We stress that Figure 9 compares the parallel efficiency of the RF and the PME methods and does not compare different MD applications. As shown in Figures 9 and 10, the use of the RF method greatly improves the strong scaling of million-atom systems, to the point where 28 ns/day are

obtained when a 5.4 million atom system is run on 30k cores with a 2 fs time step. Using RF for the electrostatics calculation removes the biggest inherent limitation of the scaling of MD. While all (i.e., irrespective of the method of treating the electrostatics) MD simulations in the NPT ensemble require one global communication (a MPI_Allreduce for a barostat and a global thermostat), this communication is not necessary at each step. The FFT part of PME, however, requires two additional global MPI_Alltoall communications, which take more time than MPI_Allreduce and do not exhibit good weak scaling, see Figure 11. Consequently, the performance, in ns/day, for large systems is inherently limited with PME.

The RF method has been employed in numerous studies,[14,22−30] and there has been some validation of its use on biological systems, including systems with net charges.[14,23−25,27] However, some of these earlier studies involved short (~5 ns) trajectories, and therefore, it cannot be guaranteed that the RF does not induce a significant bias when dealing with longer time scale dynamics. Further benchmarks are needed to examine the applicability of the RF method for biomolecular systems containing charged groups. Indeed, both the gains in computational efficiency and the possible sources of error arise from the implicit treatment of the Coulomb interaction for atoms separated more than the cutoff distance.

In the present work, the RF method was employed using an infinite dielectric constant outside the cutoff sphere, the benefit of this approach being that the Coulomb force is continuous (and equal to zero) beyond the cutoff distance. In the present study of systems that do not contain net charges, the RF does not appear to compromise the accuracy of MD simulation of the test system under study. This conclusion is drawn after a series of tests in which simulations were performed with different methods for treating the long-range electrostatics interactions. The RF and shift/switch methods are similar in the sense that they do not consider explicitly electrostatic interactions between atoms separated by more than the cutoff distance. Consequently, one might have expected the RF and shift methods to yield similar results. However, our findings suggest a different picture: all benchmarks show very good agreement between RF and PME, while the shift method exhibits several significant artifacts. Also the RF and PME simulations are in very good agreement in tests on bulk water.

In the near future, it is anticipated that the performance of MD using RF might be improved to over 60 ns/day for million-atom systems by using threads and asynchronous communication with neutral territory for improving parallel efficiency. In further benchmarks using the RF together with all-bond constraints and virtual sites, which allow removal of hydrogen atom degrees of freedom enabling integration time steps up to 5 fs,[4] we found that 38 ns/day is obtained when the 5.4-million atom system is run on 15 360 cores (data not shown). Since PME simulations are limited by the time step of the full electrostatics (e.g., 6 fs in the comparison), a longer time step for the short-range interaction does not improve the performance of PME as it would for RF. Thus for large systems, a significant improvement in

Biological Molecular Dynamics Simulation

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2807**

performance by employing longer time steps is more easily achieved using RF for the electrostatics. However, we stress that carefully designed benchmarks should first be performed before 5 fs time step simulations are routinely applied to biomolecular systems.

This work focuses on supercomputer architectures similar to the Cray XT5. In the future accelerators (GPU, Cell), special purpose and multicore clusters and MPP are likely to become competitive with general purpose supercomputers. For example, recently the special purpose computer Anton was able to simulate a 23k atom protein with a speed of over 10 μs/day[63] using Gaussian split Ewald.[64] Furthermore, accelerators have shown very good performance on MD with small numbers of atoms and processors.[65,66] For the challenging task of using accelerators in highly parallel computers in ref 66, 2 ns/day were obtained for a one million atom system.

Other possibilities for improving the parallel performance of electrostatics are to employ multigrid[64,67−69] or multipole methods.[70] These algorithms are faster for very large systems because the work scales as $O(N)$, compared to $O(N \log N)$ for PME. A larger prefactor, however, can make these methods slower for small systems. A theoretical estimate has predicted the real-space Gaussian split Ewald multigrid method to be faster than the reciprocal FFT method (very similar to PME) for systems larger than $\sim$30 000 atoms.[64] For the protein ApoA-I (with 92 224 atoms), it has been shown that the multigrid method implemented in PROTO MOL is as fast the PME in NAMD 2.5 (tested up to 66 processors).[68]

Some critical biological phenomena, such as ligand binding and the folding of small proteins, require the simulation of relatively small systems (e.g., $\sim$10^4 atoms or $\sim$1−10 nm length scales) for relatively long time scales (e.g., 10^3 s). For this type of application the strategy described here is not applicable. Rather, the present approach permits efficient atomistic MD simulation of larger, multimillion-atom biomolecular systems that do not contain net charges (i.e., on a length scale $\sim$100 nm) for times of $\sim$30 ns/day. Using the proposed strategy simulations of these large systems for time scales approaching the microsecond would now seem to be within reach on the Cray XT5. We anticipate that a wealth of structural and dynamical information of biological importance will, thus, be revealed.

Supporting Information Available: Correctness of the converted topology and force field that was checked by comparison with CHARMM and NAMD is outlined. The amplitudes of the eigenvalues obtained from the PCA of the trajectories using the three methods are presented. Weak scaling information of the complex-to-complex FFT on Cray XT5 with FFT implemented. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) *Computational Biochemistry and Biophysics*, 1st ed.; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Watanabe, M., Eds.; Marcel-Decker, Inc.: New York,2001.

(2) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(3) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(4) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(5) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.

(6) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE Conf. Supercomput.*; New York, 2006.

(7) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; Mcpherson, A.; Schulten, K. *Structure* **2006**, *14*, 437–449.

(8) Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470–480.

(9) Zink, M.; Grubmüller, H. *Biophys. J.* **2009**, *96*, 1350–1363.

(10) Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

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

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

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

(14) Baumketner, A. *J. Chem. Phys.* **2009**, *130*, 104106+.

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

(16) Villarreal, M. A.; Montich, G. G. *J. Biomol. Struct. Dyn.* **2005**, *23*, 135–142.

(17) Hünenberger, P. *Biophys. Chem.* **1999**, *78*, 69–88.

(18) Gunsteren, W. F.; Berendsen, H. J.; Rullmann, J. A. *Faraday Discuss. Chem. Soc.* **1978**, *66*, 58–70.

(19) Neumann, M. *Mol. Phys.* **1983**, *50*, 841–858.

(20) Neumann, M. *J. Chem. Phys.* **1986**, *85*, 1567–1580.

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

(22) Hunenberger, P. H.; van Gunsteren, W. F. *J. Chem. Phys.* **1998**, *108*, 6117–6134.

2808 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Schulz et al.

(23) Walser, R.; Hünenberger, P. H.; van Gunsteren, W. F. *Proteins* **2001**, *43*, 509–519.

(24) Nina, M.; Simonson, T. *J. Phys. Chem. B* **2002**, *106*, 3696–3705.

(25) Gargallo, R.; Hünenberger, P. H.; Avilés, F. X.; Oliva, B. *Protein Sci.* **2003**, *12*, 2161–2172.

(26) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847–10860.

(27) Baumketner, A.; Shea, J. E. *J. Phys. Chem. B* **2005**, *109*, 21322–21328.

(28) van der Spoel, D.; van Maaren, P. J. *J. Chem. Theory Comput.* **2006**, *2*, 1–11.

(29) Gnanakaran, S.; Nussinov, R.; Garcia, A. E. *J. Am. Chem. Soc.* **2006**, *128*, 2158–2159.

(30) Matthes, D.; de Groot, B. L. *Biophys. J.* **2009**, *97*, 599–608.

(31) Cosgrove, D. J. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 850–861.

(32) Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.

(33) Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236–1243.

(34) Petridis, L.; Smith, J. C. *J. Comput. Chem.* **2009**, *30*, 457–467.

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

(36) van der Spoel, D.; Lindahl, E.; Hess, B.; Kutzner, C.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. *GROMACS USER MANUAL*, Version 4.0; www. GROMACS.org.

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

(38) Nishiyama, Y.; Langan, P.; Chanzy, H. *J. Am. Chem. Soc.* **2002**, *124*, 9074–9082.

(39) Ding, S. Y.; Himmel, M. E. *J. Agric. Food. Chem.* **2006**, *54*, 597–606.

(40) Petridis, L.; Xu, J.; Crowley, M. F.; Smith, J. C.; Cheng, X. Atomistic Simulation of Lignocellulosic Biomass and Associated Cellulosomal Protein Complexes. In *Computational Modeling in Lignocellulosic Biofuel Production*; Nimlos, M. R., Crowley, M. F., Eds.; ACS, 2009; p in print.

(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(42) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

(43) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695+.

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

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

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

(47) McLain, S.; Soper, A.; Daidone, I.; Smith, J.; Watts, A. *Angew. Chem., Int. Ed.* **2008**, *47*, 9059–9062.

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

(49) Gullingsrud, J.; Saam, J.; Phillips, J. psfgen; http://www. ks.uiuc.edu/Research/vmd/plugins/psfgen/ 2006 (accessed January 6, 2009).

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

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

(52) Gullingsrud, J. catdcd; http://www.ks.uiuc.edu/Development/ MDTools/catdcd/ (accessed January 6, 2009).

(53) Tarini, M.; Cignoni, P.; Montani, C. *IEEE Trans. Visual Comput. Graphics* **2006**, *12*, 1237–1244.

(54) Oster, G.; Kirkwood, J. G. *J. Chem. Phys.* **1943**, *11*, 175–178.

(55) Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–152.

(56) French, A. D.; Johnson, G. P. *Cellulose* **2004**, *11*, 449–462.

(57) Schulten, K.; Phillips, J. C.; Kal Bhatele., A. Biomolecular modeling in the era of petascale computing. In *Petascale Computing: Algorithms and Applications*; Bader, D., Ed.; Chapman and Hall/CRC Press, Taylor and Francis Group: Boca Raton, FL, 2008; pp 165−181.

(58) Plimpton, S. Parallel FFT Package; http://www.sandia.gov/ ∼ sjplimp/docs/fft/README.html (accessed January 6, 2009).

(59) Takahashi, D. FFTE: A Fast Fourier Transform Package; http://www.ffte.jp (accessed January 6, 2009).

(60) Frigo, M.; Johnson, S. *Proc. IEEE* **2005**, *93*, 216–231.

(61) Fitch, B. G.; Rayshubskiy, A.; Eleftheriou, M.; Ward, C. T. J.; Giampapa, M.; Pitman, M. C.; Germain, R. S. Blue matter: approaching the limits of concurrency for classical molecular dynamics. *Proc. ACM/IEEE Conf. Supercomput.*; New York, 2006.

(62) Freddolino, P. L.; Liu, F.; Gruebele, M. H.; Schulten, K. *Biophys. J.* **2008**, *94*, 75−77.

(63) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(64) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *J. Chem. Phys.* **2005**, *122*, 054101.

(65) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.

(66) Phillips, J. C.; Stone, J. E.; Schulten, K. Adapting a message-driven parallel application to GPU-accelerated clusters. *Proc. ACM/IEEE Conf. Supercomput.*; Piscataway, NJ, 2008; pp 1−9.

(67) Sagui, C.; Darden, T. *J. Chem. Phys.* **2001**, *114*, 6578–6591.

(68) Izaguirre, J. A.; Hampton, S. S.; Matthey, T. *J. Parallel Distr. Com.* **2005**, *65*, 949–962.

(69) Skeel, R. D.; Tezcan, I.; Hardy, D. J. *J. Comput. Chem.* **2002**, *23*, 673–684.

(70) Kurzak, J.; Pettitt, B. M. *Mol. Simulat.* **2006**, *32*, 775–790.

JCTC Journal of Chemical Theory and Computation

Polarizable Force Field for Protein with Charge Response Kernel

Miho Isegawa and Shigeki Kato*

Department of Chemistry, Graduate School of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan

Abstract: We present a molecular mechanical force field for polypeptides and proteins involving the electronic polarization effect described with the charge response kernel. All of the electrostatic parameters for 20 amino acids are obtained by ab initio electronic structure calculations and combined with the AMBER99 force field. The refittings of dihedral angle parameters in the torsional potentials are performed so as to reproduce the ab initio optimized geometries and relative energies for the conformers of dipeptides. The present force field is applied to molecular dynamics simulation calculations of the extended alanine tetra and cyclic pentapeptides in aqueous solution. The infrared spectra are calculated in order to analyze the charge polarization effect on the spectral profiles.

1. Introduction

Molecular dynamics (MD) simulation is a powerful means for studying statistical and dynamic properties of biological systems. In these studies, the reliability of calculated results is governed by the quality of employed molecular mechanical (MM) force fields. The standard MM force fields such as AMBER,[1−3] CHARMM,[4−6] GROMOS,[7−9] and OPLS[10−12] have been developed and widely used for MD simulations of proteins.

In order to construct more realistic force fields, the charge polarizable models have been proposed in recent years[13−22] and applied to various biological systems.[23−27] The charge polarization induced by intra- or intermolecular interaction has been represented by the site point dipole,[13−15] fluctuating charge,[16−18] and Drude oscillator[19−21] methods. Thus, the AMBER, CHARMM, and OPLS force fields have been refined with combining the polarizable electrostatic interaction.[14,15,17,21]

In the present work, we develop a charge-polarizable MM model for peptides and proteins based on the charge response kernel (CRK) method.[28−31] The advantage of the CRK model is to be able to describe the nonlocal charge polarization through the off-diagonal elements, which is important for representing the charge migration through peptide bonds and aromatic residues. It is also noted that the CRK matrix elements can be directly determined by ab inito electronic structure calculations for isolated molecules. We employ the second-order Møller−Plesset perturbation (MP2) method for calculating the CRK matrices in this study.[31]

The CRK model has been applied to water, aprotic solvents, and organic radicals, and the infrared (IR) and Raman spectra as well as the diffusion constants in the liquid phase have been calculated.[31−33] Although the applications of the CRK model have been limited to small molecules so far, we extend it for use with polypeptides and proteins in the present study. For this purpose, we devise a method to construct the CRK matrix of protein using the matrices of fragment amino acids.

We combine the present CRK model with the AMBER99[3] potential functions for bond stretching, bending, and torsional and nonbonding van der Waals interactions. As is known, the torsional potentials are strongly correlated to the intramolecular electrostatic interaction and play an important role in determining the conformational stability.[34,35] We therefore readjust the parameters of torsional potentials so as to reproduce the ab initio geometries and relative energies for various conformers of dipeptides. Such a procedure was already employed by Kaminski et al.[34] for improving the OPLS-AA force field, which shows good performance for 20 amino acid dipeptides, and by Wang et al.[35] in construct-

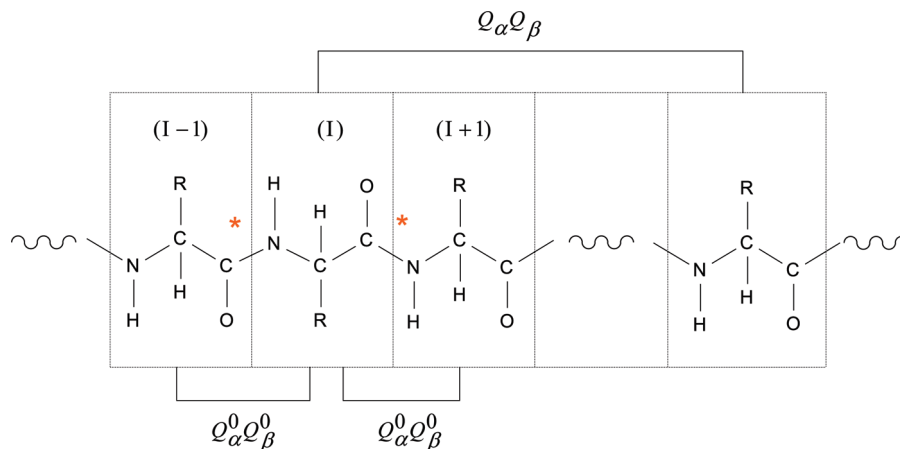* Corresponding author e-mail: shigeki@kuchem.kyoto-u.ac.jp.

Figure 1. Illustration of electrostatic interactions. The interactions are classified into two patterns: interactions by the reference charges in the nearest neighbor units and (ii) interactions, except for case i, by the polarized charges.

ing a polarizable AMBER force field for alanine dipeptide based on the site point dipole model.

To assess the utility of the present CRK-based MM model, we carry out MD calculations for the extended alanine tetrapeptide and cyclic pentapeptide in aqueous solution. Alanine is the most representative amino acid because of its simplicity and common characteristics with 20 amino acids. Actually, alanine polypeptides have been the subject of many theoretical studies, including the electronic structure and MD simulation methods.[36−41] We calculate the IR spectra of the extended and cyclic peptides because the charge distribution induced by the intra- and intermolecular electrostatic potential is expected to be different between the two forms. The IR spectra for amide modes have been investigated, and the effect of hydrogen bonding on the spectral profiles has been examined.[42−48] Bouř and Keidering[47] carried out MD simulations to calculate the IR intensity of the amide I band for linear alanine pentapeptide in aqueous solution. More recently, Torii[48] calculated the IR and Raman spectra of tetrapeptide in aqueous solution and analyzed the profile of the amide I band in detail. Here, we examine the polarization effects on the calculated spectra by decomposing the spectral shape into the contributions from the intra- and intermolecular electrostatic interaction.

In the next section, we construct the CRK-based polarizable MM model for peptides and proteins combined with the AMBER force field. The electrostatic parameters are determined for 20 amino acids and terminal species, referring to the results of ab initio electronic structure calculations. We optimize the Fourier coefficients in dihedral angle terms so as to reproduce the ab initio relative energies and geometries for the conformers of dipeptides. The results are presented for alanine di- and tetrapeptide in section 2, and for the other dipeptides in the Supporting Information. In section 3, we apply the present force field to MD simulations of the extended alanine tetrapeptide and cyclic pentapeptide in aqueous solution. The effect of intra- and intermolecular hydrogen bonding on the spectral profiles is discussed. The concluding remarks of the present work are summarized in section 4.

2. Force Field

2.1. Electrostatic Potential Function. In the present polarizable force field for polypeptide with the CRK model, the partial charge $Q_{\alpha I}$ at the atomic site α in the Ith amino acid unit is defined by

$$Q_{\alpha I} = Q_{\alpha I}^0 + \sum_{\beta \in J}^{N_J} \sum_{J}^{N_u} K_{\alpha I, \beta J} V_{\beta J}^{\text{eff}} \tag{1}$$

where N_u is the number of the amino acid units in the polypeptide and N_J is the number of the atomic sites in the unit J. $Q_{\alpha I}^0$ is the reference charge initially assigned on the atomic site (see Figure 1). The induced charge is given by the second term of eq 1, where $K_{\alpha I, \beta J}$ is the CRK matrix element and $V_{\beta J}^{\text{eff}}$ is the effective electrostatic potential acting on the β site in the Jth unit. Note that the CRK matrix is related to the polarizability as

$$\alpha_{mn} = -\sum_{\alpha \in I}^{N_I} \sum_{I}^{N_u} \sum_{\beta \in J}^{N_J} \sum_{J}^{N_u} K_{\alpha I, \beta J} R_{m, \alpha I} R_{n, \beta J} \tag{2}$$

where $R_{m, \alpha I}$ is the Cartesian coordinate of the atomic site α in the unit I with $m = x, y, z$.

In the conventional CRK approach, the reference charge $Q_{\alpha I}^0$ is defined to be the atomic charge of isolated molecule, and only the intermolecular contribution to the electrostatic potential is taken into account for calculating $V_{\beta J}^{\text{eff}}$. However, it is impractical to employ such a method to represent the atomic site charges of the whole polypeptide because the charges are affected by the intramolecular hydrogen bonding and thus strongly depend on the conformation of the polypeptide. In order to avoid such a difficulty, we obtained the atomic charges for each amino acid unit in the isolated condition and assigned them to the reference charges. The charge redistribution due to the interaction between different amino acid units is thus given by the second term of eq 1.

The CRK matrix of the whole polypeptide was also constructed from the matrices obtained for the constituent amino acids. The CRK matrix for glycine tripeptide was used to represent the matrix elements corresponding to the amide bond. Since the CRK matrix must have one zero eigenvalue

with the eigenvector corresponding to the uniform potential, we applied the projection operator $\mathbf{P} = \mathbf{I} - 1/N_a\mathbf{1}$, with N_a being the number of total atomic sites in the polypeptide, as

$$\mathbf{K} = \mathbf{P}\widetilde{\mathbf{K}}\mathbf{P}^t \qquad (3)$$

where $\widetilde{\mathbf{K}}$ is the CRK matrix constructed from those of the fragments. Here, $\mathbf{I}$ is the unit matrix and $\mathbf{1}$ is the matrix with all elements of 1. With this operation, the total charge of the polypeptide is conserved.

The effective electrostatic potential acting on the β site in the Jth unit has both the intra- and intermolecular contributions as mentioned above. The intramolecular part is given by

$$V_{\beta J}^{\text{eff}} = \sum_{\gamma \in L}^{N_L} \sum_{L}'' Q_{\gamma L} \frac{f(|\mathbf{R}_{\beta J} - \mathbf{R}_{\gamma L}|)}{|\mathbf{R}_{\beta J} - \mathbf{R}_{\gamma L}|} \qquad (4)$$

where the index L runs over all of the amino acid units except for J and $J \pm 1$, and $f(|\mathbf{R}_{\beta J} - \mathbf{R}_{\gamma L}|)$ is the damping function introduced to attenuate the electrostatic potential at the short range. We omitted the electrostatic contributions from the sites in the nearest neighbor units because the sites in more distant units become important in characterizing the conformation of polypeptide. The site charges are determined so as to satisfy eqs 1 and 4 simultaneously.

Using the intramolecular electrostatic potential defined by eq 4, the electrostatic interaction part of the potential function for an isolated polypeptide is written as

$$
\begin{aligned}
U_{\text{ES}} = {} & \frac{1}{2} \sum_{\alpha \in I}^{N_I} \sum_{I}^{N_u} \sum_{\beta \in J}^{N_J} \sum_{J}' Q_{\alpha I}^0 Q_{\beta J}^0 \frac{f_{\alpha I, \beta J}^0}{|\mathbf{R}_{\alpha I} - \mathbf{R}_{\beta J}|} \\
& + \frac{1}{2} \sum_{\alpha \in I}^{N_I} \sum_{I}^{N_u} \sum_{\beta \in J}^{N_J} \sum_{J}'' Q_{\alpha I} Q_{\beta J} \frac{f(|\mathbf{R}_{\alpha I} - \mathbf{R}_{\beta J}|)}{|\mathbf{R}_{\alpha I} - \mathbf{R}_{\beta J}|} \\
& - \frac{1}{2} \sum_{\alpha \in I}^{N_I} \sum_{I}^{N_u} \sum_{\beta \in J}^{N_J} \sum_{J}^{N_u} K_{\alpha I, \beta J} V_{\alpha I}^{\text{eff}} V_{\beta J}^{\text{eff}}
\end{aligned} \qquad (5)
$$

where the first term represents the interactions between the reference charges on the nearest neighboring units. The factor $f_{\alpha I, \beta J}^0$ is introduced to follow the interaction scheme in AMBER; the 1−2 and 1−3 nonbonding interactions are omitted, and the 1−4 interaction is estimated by multiplying the factor of 1/1.2. The second term includes all of the interactions between the amino acid units except for the nearest neighbors. Note that the charge polarization effect due to the intramolecular electrostatic interaction is incorporated in this term. Finally, the third term corresponds to the electronic reorganization energy originated from the charge polarization in the polypeptide.

We employed Thole's damping function[49] for $f(R)$ in eqs 4 and 5,

$$
f(R) = \begin{cases} \left(\frac{R}{S}\right)^4 - 2\left(\frac{R}{S}\right)^3 + 2\left(\frac{R}{S}\right) & (R < S) \\ 1 & (R > S) \end{cases} \qquad (6)
$$

with

$$S = A(\alpha_\alpha \alpha_\beta)^{1/6} \qquad (7)$$

Here, R is the distance between the atomic sites α and β, and the critical distance S is determined by the factor A and the atomic polarizability α_α. We set the factor A to be 2.8,[31,33] which was used for the intermolecular interaction in liquid water and aprotic solvent systems, and the polarizalibities as $\alpha_H = 0.514$, $\alpha_C = 1.405$, $\alpha_{CA} = 1.671$, $\alpha_{CT} = 1.822$, $\alpha_N = 1.105$, $\alpha_{NA} = 1.296$, $\alpha_O = 0.862$, and $\alpha_S = 2.540$ in Å^3.[49,50] Here, the subscripts CA, CT, and NA mean the carbonyl carbon, sp^3 carbon, and amide nitrogen, respectively.

2.2. Electrostatic Parametrization. We carried out ab initio calculations for 20 neutral and five charged amino acids at the MP2 level of theory with the 6-31G** basis set. The C-terminal N-methyl group (NME) and N-terminal acetyl group (ACE) were also considered. In the calculations, both the terminals of amino acid were blocked with methyl groups, as shown in Figure 1 with the asterisk. However, the methyl groups were replaced with the hydrogen atoms for relatively large amino acids (i.e., arginine, lysine, phenylalanine, tyrosine, and tryptophan) in order to reduce the calculation cost.

We first performed the geometry optimizations for all of the species at the MP2 level with the 6-31G** basis set. At these optimized geometries, the electrostatic potential (ESP) charges on the atomic sites were obtained by[31]

$$Q_\alpha = \frac{\partial E^{\text{MP2}}}{\partial V_\alpha} \qquad (8)$$

where E^{MP2} is the MP2 energy and V_α is the electrostatic potential acting on the site α. We modified the ESP charges by imposing the following three conditions. First, the site charges such as on the hydrogens of the methyl group and the symmetrical carbons and hydrogens in the aromatic ring were replaced with the averaged values. Second, the charges in the amide part were taken from the averaged values of glycine dipeptide at three stable conformers, C7, C5, and α:[51] −0.473 for N, 0.311 for H, 0.580 for C, and −0.500 for O. Finally, the charges of atoms in the side chain were scaled so that the value of the total charge becomes an integer, that is, 0 for neutral and ±1 for charged amino acids, under the condition that the charges of N, H, C, and O in the amide part are fixed. The parameters for the terminal capping groups, NME and ACE, were also determined by taking the average of the charges at three conformers of glycine dipeptide as in the case of the amide part. The reference charges thus obtained are summarized in Tables S1−S27 in the Supporting Information.

We also executed ab initio calculations to obtain the CRK matrices of all of the species defined by[31]

$$K_{\alpha\beta} = \frac{\partial^2 E^{\text{MP2}}}{\partial V_\alpha \partial V_\beta} \qquad (9)$$

The CRK matrix has large negative values at the diagonal elements, and the magnitudes of off-diagonal elements rapidly decrease with increasing the number of bonds separating the two sites. Thus, the 1−4 elements become negligibly smaller compared to the diagonal ones. We therefore truncated the CRK matrices at the 1−3 elements to reduce the conformational dependency. It is however noted
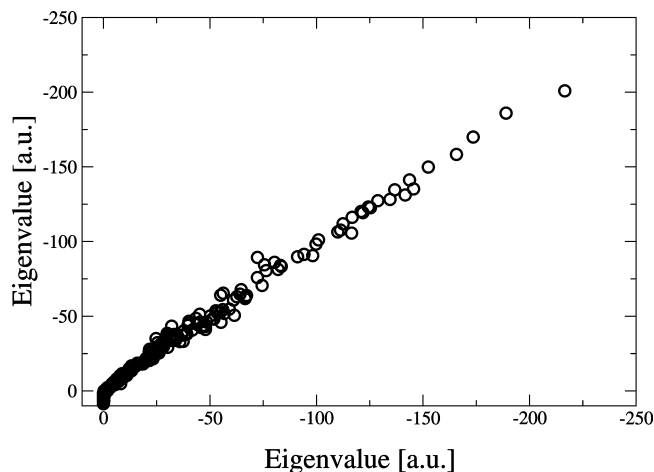
Figure 2. Correlation of eigenvalues between original and truncated CRK matrices for glycine dipeptide and 20 amino acids. Horizontal and vertical axes are for original and truncated matrices.

that the elements beyond the 1−3 of the aromatic ring are not small because of π conjugation and cannot be ignored. For this reason, we retained all of the elements in the aromatic ring for phenylalanine and tyrosine, as well as the elements between the α carbon and aromatic sites for tryptophan. All of the CRK matrices are listed in Tables S1−S27 (Supporting Information). Figure 2 shows the eigenvalues of truncated CRK matrices compared with the original eigenvalues for a glycine dipeptide, and 19 neutral and five charged amino acids, where we can see that the truncated eigenvalues are very close to the original ones.

2.3. Parametrization of Dihedral Potential. The intramolecular potential function for isolated polypeptide is the same as the AMBER force field except for the electrostatic term U_{ES},

$$U = \sum_{\text{bonds}} K_r(R - R_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 +$$

$$\sum_{\text{dihedrals}} V_n[1 + \cos(n\phi - \gamma)] + \sum_{\alpha > \beta} \left[\frac{A_{\alpha\beta}}{R_{\alpha\beta}^{12}} - \frac{B_{\alpha\beta}}{R_{\alpha\beta}^6} \right] + U_{ES}$$

$$(10)$$

Here, the first two terms represent the harmonic bond stretching and bending motions, respectively. The third term is the torsional potential written by the Fourier series. The last two terms are the nonbonding van der Waals and electrostatic interactions.

Since the stretching and bending terms are rather independent of the electrostatic term, we adopted the same parameters as those for AMBER99 for these terms. However, the dihedral terms strongly correlate to the electrostatic term and are responsible to the conformational stability. We therefore reparametrized the torsion parameters V_n so as to reproduce the relative energies and geometries of several stable dipeptide conformers obtained with the ab inito electronic structure calculations. In order to determine the parameters for the backbone dihedral angles (ϕ, ψ), alanine dipeptide was selected because the effect of the side chain on the backbone conformation is small. We generated 1296
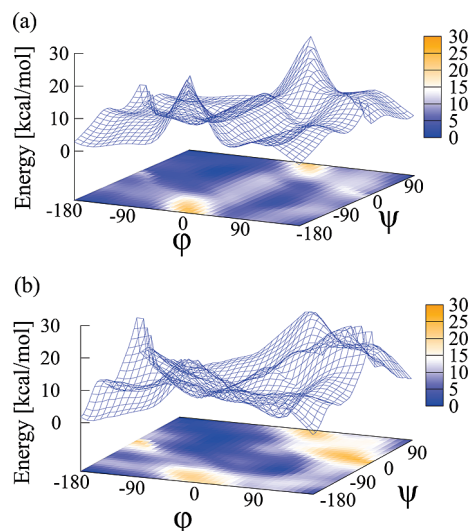

Figure 3. Potential energy surfaces as a function of (ϕ, ψ) obtained by (a) ab initio calculation and (b) the CRK model. The most stable conformer C7$_{eq}$ is taken to be 0 kcal/mol.

reference grid points in total (36 points are taken for each dihedral angle) and performed ab initio calculations at the MP2 level with the 6-31G** basis set. The ab initio calculated quantum mechanical (QM) potential surface is shown in Figure 3a as a function of the dihedral angles (ϕ, ψ). The Fourier coefficients of torsional potential are obtained by minimizing the function:

$$R = \sum_i^{\text{ngrid}} \omega_i(\Delta E_i^{\text{MM}} - \Delta E_i^{\text{QM}})^2 \qquad (11)$$

where the index i refers a grid point and ω_i is the weighting factor

$$\omega_i = \exp(-\zeta \Delta E_i^{\text{QM}}) \qquad (12)$$

with

$$\Delta E_i^{\text{QM}} = E_i^{\text{QM}} - E_{\text{C7}_{eq}}^{\text{QM}} \qquad (13)$$

and

$$\Delta E_i^{\text{MM}} = (E_i^{\text{MM}} - E_i^{\text{torsion}}) - (E_{\text{C7}_{eq}}^{\text{MM}} - E_{\text{C7}_{eq}}^{\text{torsion}}) \qquad (14)$$

Here, E^{torsion} is the torsion energy corresponding to the backbone dihedral angles ϕ and ψ, and the subscript C7$_{eq}$ is the conformer with the angles ϕ and $\psi = -85.8°$ and 79.2°, which is the most stable conformer of dipeptide with the intramolecular hydrogen bond. The parameter ζ in the weighting factor is determined so as to minimize the relative energy rms deviation from the ab initio reference values for the five stable conformers: C7$_{eq}$, C5, C7$_{ax}$, α_L, and α'.

The optimized torsional parameters of the main chain determined for alanine dipeptide were used for most of the other dipeptides. However, when either the relative energy or the optimized geometry extremely differs from the ab initio reference, the least-squares fittings of dihedral angle parameters for the side chain were performed. In this fitting,

ab initio grid points for the side chain are generated in a crosslike manner for each stable conformer.[34]

2.4. Evaluation of the Force Field for Isolated Peptide. The fitted dihedral parameters are summarized in Table S28 in the Supporting Information. To evaluate the present polarizable MM force field, we compared the relative energies and key dihedral angles at the stable conformations of all of the dipeptides with those obtained by ab initio calculations. For this purpose, we carried out the geometry optimizations of the stable conformers for 20 neutral and five charged dipeptides at the HF level with the 6-31G** basis set, starting from the initial geometries given by Kaminski et al.[34] At the optimized geometries, the MP2 calculations were repeated to estimate the relative energies. The resultant dihedral angles are summarized in Tables S29−S46 (Supporting Information).

We next carried out the geometry optimizations of the conformers for each peptide with the present polarizable MM model, where the ab initio optimized geometries were used as the initial guess. The calculations were also performed with the original AMBER99 force field. Since there are several stable conformers for dipeptides, the rms deviation from the ab initio reference energies was calculated by

$$R_E = \sqrt{\frac{1}{N_C} \sum_i^{N_C} (E_i^{MM} - E_i^{QM})^2} \qquad (15)$$

where N_C is the number of conformers and E_i^{MM} and E_i^{QM} correspond to the MM and ab initio energies, respectively. The rms deviation of the dihedral angles was also given by

$$R_D = \sqrt{\frac{1}{N_C N_D} \sum_i^{N_C} \sum_j^{N_D} (\Psi_{ij}^{MM} - \Psi_{ij}^{QM})^2} \qquad (16)$$

where N_D is the number of key dihedral angles for a peptide and Ψ_{ij}^{MM} and Ψ_{ij}^{QM} are the dihedral angles at the MM and ab initio optimized geometries, respectively. It is noted that there were conformers whose stable geometries show very large deviations from the corresponding ab initio references for some peptides. For such a case, we excluded the results of these conformers from the estimations of rms deviations.

Alanine Dipeptide and Tetrapeptide. Figure 3b shows the potential energy surface as a function of ϕ and ψ for alanine dipeptide and its contour map calculated with the present MM model. We used the same geometries as for the calculation of the ab initio surface, Figure 3a. Compared with the ab initio surface, we can see that the low potential energy regions involving the conformers C7$_{eq}$, C5, and C7$_{ax}$ are well reproduced, while the high-energy regions, $(-180° \leq \phi \leq -140°, -30° \leq \psi \leq 40°)$ and $(50° \leq \phi \leq 180°, -10° \leq \psi \leq 100°)$, are too high. This is because we employed the weighting factor, which emphasizes low-energy grid points in determining the torsional potential parameters. Table 1 shows the rms deviations of relative energies and dihedral angles for alanine dipeptide. The zero of ab initio relative energy was that of the most stable conformer C7$_{eq}$, while the MM energies were shifted so as to minimize the rms deviation. The five conformers, C7$_{eq}$, C5, C7$_{ax}$, α_L, and α', are reproduced well, but there is no minimum corresponding

Table 1. Alanine Dipeptide Relative Energies [kcal/mol] and RMS Deviations of Backbone Dihedral Angles [deg] from the ab Initio References[a]

conformer	ab initio[b]	(ϕ, ψ)	CRK	AMBER
C7$_{eq}$	0.00	(−85.7, 79.1)	−0.03/12.4	0.25/29.5
C5	1.33	(−157.2, 159.8)	1.31/14.1	1.15/5.9
C7$_{ax}$	2.38	(76.0, −55.4)	2.32/15.4	2.48/20.9
β_2	3.00	(−130.9, 22.3)		2.83/1.7
α_L	4.42	(67.0, 30.2)	4.51/30.8	
α'	5.98	(−166.0, −39.9)	5.99/22.6	
rmsd			0.05/20.2	0.18/18.3

[a] RMS deviations are in the bottom. [b] MP2 6-31G**//HF6-31G**.

Table 2. Alanine Tetrapeptide Relative Energies [kcal/mol] and RMS Deviations of Backbone Dihedral Angles [deg] from the ab Inito References[a]

conformer	ab inito[b]	CRK	AMBER
1	5.64	4.67/13.6	4.92/6.3
2	5.28	4.58/15.2	5.16/18.0
3	0.00	1.04/12.4	2.93/18.6
4	6.77	4.45/15.9	
5	7.18	4.98/14.7	4.41/8.1
6	2.39	4.63/23.4	2.25/22.3
7	6.45	6.21/9.9	
8	4.63	7.01/16.2	7.53/19.6
9	7.91	8.65/17.2	7.71/23.9
10	7.40		5.51/30.1
rmsd		1.63/15.8	1.90/19.8

[a] RMS deviations are in the bottom. [b] MP2 6-31G**//HF6-31G**.

to β_2 on the MM potential surface. This is because the β_2 minimum exists at the high-energy regions in Figure 3, and the shallow minimum corresponding to this conformer disappears on the potential surface with the present force field. As seen in Table 1, the original AMBER force field reproduces the four conformers C7$_{eq}$, C5, C7$_{ax}$, and β_2, but the conformers α_L and α' are not reproduced. Wang et al.[35] optimized the atomic charges and dihedral parameters for alanine dipeptide taking into account the polarization effect with the AMBER02 force field. However, the optimized force field failed to reproduce some of the stable conformers. Those results indicate that we may need to use a more elaborate form of the potential function including the coupling terms between the torsional angles in order to reproduce all of the conformers.

The transferability of the potential parameters determined for dipeptides to larger polypeptides is crucial in constructing a MM force field for proteins. We applied the present parameters to alanine tetrapeptide. Table 2 shows the relative energies at the optimized geometries and the rms deviations of dihedral angles for 10 conformers[52] obtained with the present and AMBER force fields. The numbering of the conformers was taken from ref 52. Conformer **3** is the most stable with the ab initio and present force fields because of the existence of intramolecular hydrogen bonding, while conformer **6** is more stable than conformer **3** with the AMBER force field. Conformer **10** with high energy was not reproduced with the present model. The angles ϕ_1 and ψ_1 at the converged geometry largely deviate from the ab initio values. On the other hand, the AMBER force field failed to reproduce the ab initio geometries of conformers **4** and **7**.

Asparagine and Glutamine Dipeptides. Both of the peptides have an amide group in the side chain which participates in the formation of intramoleular hydrogen bonds. Actually, the most stable ab initio conformer of asparagine is stabilized by the two intramolecular hydrogen bonds between the atoms in the backbone and side chain with bond distances of 2.34 and 2.14 Å, respectively. For the asparagine dipeptide, the ab initio stable conformers are reproduced well by the present model with the relative energy and dihedral rms deviations, 0.43 kcal/mol and 15.8°, respectively. For the glutamic acid, two of the 11 conformers obtained by the ab initio calculations are not reproduced with both the present and AMBER force fields. Note that we used the original AMBER parameters for the dihedral angles for the side chains for both the peptides. The results of the two peptides are shown in Tables S47 and S48 (Supporting Information).

Valine, Isoleucine, and Leucine Dipeptides. The side chains of these peptides are composed of alkyl groups. For the valine and isoleuicine dipeptides, all of the ab initio conformers are well reproduced. The resultant rms deviations of both the peptides are ∼0.4 kcal/mol for the relative energies and ∼9.0° for the dihedral angles, as shown in Tables S49 and S50 (Supporting Information). However, there are large deviations of dihedral angles of the main chain from the ab initio values for conformers **3**, **6**, and **7** of leucine, as shown in Table S51 (Supporting Information).

Serine and Threonine Dipeptides. A hydroxyl group is included in the side chains of both the peptides. For the serine dipeptide, we refitted the dihedral parameters for the side chain, χ_1 (N−CT−CT−OH), χ'_1 (OH−CT−CT−C), and χ_2 (CT−CT−OH−HO), keeping the main chain torsion parameters fixed. The rms deviation of 1.55 kcal/mol for the relative energy and a value of 9.5° for the dihedral angle are comparable with those of the AMBER force field, 1.72 kcal/mol and 8.1°, respectively, as shown in Table S52 (Supporting Information). Note that, for some conformers of threonine, the present ab inito optimized dihedral angles differ from those of the corresponding conformers in ref 34. For both force fields, the optimization of conformer **2** resulted in conformer **3**, whose dihedral angles are close to those of conformer **2** except for χ_2. The resultant rms deviations are smaller than the AMBER for both the relative energies and dihedral angles, as shown in Table S53 (Supporting Information).

Phenylalanine, Tyrosine, and Tryptophan Dipeptides. The side chains of these peptides involve an aromatic ring. The rms deviations of relative energies and dihedral angles are small for the phenylalanine and tyrosine dipeptides, as seen in Tables S54 and S55 (Supporting Information). For the tryptophan dipeptide, we executed the fittings of dihedral angle parameters for the side chain, χ_1 (N−CT−CT−C*) and χ'_1 (C*−CT−CT−C). Although two of the nine conformers with high energies could not be described with the present force field, the present model gives smaller rms deviations than the AMBER force field, as shown in Table S56 (Supporting Information).

Cysteine and Methionine Dipeptides. These three peptides involve a sulfur atom in the side chains. For the cysteine dipeptide, the refitting of dihedral parameters for the side

chains, χ_1 (N−CT−CT−SH), χ'_1 (SH−CT−CT−C), and χ_2 (CT−CT−SH−HS), was performed. As shown in Table S57 (Supporting Information), the highest-energy conformer among the five stable conformers is not reproduced, but the remaining four conformers are well reproduced with a rms deviation of 0.60 kcal/mol for the relative energy and 14.6° for the dihedral angle. For the methionine, the dihedral refittings for the side chain were not required because the ordering of the relative energy is consistent with the ab initio one with relative energy and dihedral angle rms deviations of 0.82 kcal/mol and 11.0°, respectively. The results are given in Table S58 (Supporting Information).

Charged Asparatic Acid, Glutamic Acid, Histidine, Arginine, and Lysine Diepetides. The asparatic and glutamic acids are negatively charged, while the histidine, arginine, and lysine are positively charged. For these charged peptides, since the geometries of stable conformers in solution deviate from those in the gas phase, we carried out geometry optimization in solution, as performed by Kaminski et al.[34] and used the gas-phase energies at these geometries for the fitting of parameters of the torsional potentials. For the arginine, the dihedral fitting for the side chain, χ_1 (N−CT−CT−CT) and χ'_1 (CT−CT−CT−C), was carried out, and for the other peptides, the dihedral parameters for the side chain were taken from the AMBER force field. The ordering of relative energies is reproduced well except for histidine, and the rms deviations are comparable with those of the AMBER force field. The results are shown in Tables S59−S63 (Supporting Information).

3. Application to Alanine Polypeptides

3.1. MD Simulation Method. The potential energy function of the solvated polypeptide system is expressed as

$$
\begin{aligned}
U^{\text{sol}} = U + U' + \sum_{i>j}^{N_\text{v}} \sum_{a,b}^{N_\text{s}} \left[u_{\text{LJ}}(|\mathbf{r}_{ai} - \mathbf{r}_{bj}|) + Q_{ai}Q_{bj}\frac{f(|\mathbf{r}_{ai} - \mathbf{r}_{bj}|)}{|\mathbf{r}_{ai} - \mathbf{r}_{bj}|} \right] \\
+ \sum_{i}^{N_\text{v}} \sum_{a}^{N_\text{s}} \sum_{\alpha \in I}^{N_I} \sum_{I}^{N_\text{u}} \left[u_{\text{LJ}}(|\mathbf{r}_{ai} - \mathbf{R}_{\alpha I}|) + Q_{ai}Q_{\alpha I}\frac{f(|\mathbf{r}_{ai} - \mathbf{R}_{\alpha I}|)}{|\mathbf{r}_{ai} - \mathbf{R}_{\alpha I}|} \right] \\
- \frac{1}{2} \sum_{i}^{N_\text{v}} \sum_{a,b}^{N_\text{s}} K_{ab} V_{ai} V_{bi}
\end{aligned}
$$

(17)

where N_v is the number of solvent water molecules with N_s being the number of sites in a solvent molecule. The functions U and U' are the intramolecular potential functions for the peptide, eq 10, and solvent, respectively. $Q_{\alpha I}$ includes the induced charge due to the intra- and intermolecular interactions. The electrostatic potential in the solvent reorganization energy term is written as

$$
V_{ai} = \sum_{\beta \in I}^{N_I} \sum_{I}^{N_\text{u}} Q_{\beta I}\frac{f(|\mathbf{R}_{\beta I} - \mathbf{r}_{ai}|)}{|\mathbf{R}_{\beta I} - \mathbf{r}_{ai}|} + \sum_{j(\neq i)}^{N_\text{v}} \sum_{b}^{N_\text{s}} Q_{bj}\frac{f(|\mathbf{r}_{bj} - \mathbf{r}_{ai}|)}{|\mathbf{r}_{bj} - \mathbf{r}_{ai}|}
$$

(18)

where the first and second terms are the contributions from the solute and solvent, respectively.

The electrostatic parameters for a water molecule, that is, the atomic site charges and CRK matrix elements, were

calculated at the MP2 level with the 6-31G(2d,2p) basis set, and the results are $Q_O = -0.680$ and $Q_H = 0.340$ for the site charges and $K_{OO} = -4.760$, $K_{HH} = -2.122$, $K_{OH} = 2.380$, and $K_{HH'} = -0.2580$ in atomic units for the CRK matrix elements, respectively. The Lennard-Jones (LJ) parameters, ε and σ, were taken from the SPC model,[53] and the standard combination rule is applied for constructing the site−site interaction. The intramolecular motion of water is described by the Morse potential for the stretching mode and the harmonic potential for the bending one

$$U' = D_e[1 - e^{-a_e(r-r_{eq})}]^2 + D_e[1 - e^{-a_e(r'-r'_{eq})}]^2$$
$$+ f_r(r - r_{eq})(r' - r'_{eq}) + \frac{1}{2}f_\theta r_{eq}^2(\theta - \theta_{eq})^2 \tag{19}$$

where $r_{eq} = 0.958$ Å and $\theta_{eq} = 102.9°$, the ab initio optimized values. The other parameters, D_e, a_e, f_r, and f_θ, are the same as in the previous study of water with the CRK model.[33]

For the solvent−solvent LJ terms in eq 17, the long-range part was cut off with the following function:

$$s(x) = \begin{cases} 1 & (r_{ij} < 0.9r_c) \\ 1 - 10x^3 + 15x^4 - 6x^5 & (0.9r_c < r_{ij} < r_c) \\ 0 & (r_c < r_{ij}) \end{cases} \tag{20}$$

where

$$x = (r_{ij} - 0.9r_c)/0.1r_c$$

Here, r_c is the cutoff length that is half of the box size and r_{ij} is the distance between the i and jth molecules. The electrostatic interactions between the solvent−solvent and solute−solvent are attenuated at a short range with the damping function given in eq 7, as in the case of the intramolecular interaction in the polypeptide. The Ewald sum method was employed to treat the long-range electrostatic interaction.

MD simulations of solvated alanine polypeptide systems were performed with one solute and 472 solvents for the tetrapeptide (471 solvents for the pentapeptide) in a cubic box with a length of 24.66 Å, and the periodic boundary condition was employed. The initial structures are conformer **1** in Table 2 for the tetrapeptide and the ab initio optimized structure for the pentapeptide (see Figure 4). The equations of motion were integrated using the Verlet algorithm with a time step of 0.5 fs. The charges on solute and solvent atomic sites are determined self-consistently at each time step. The equilibration of the system at 298 K was achieved by MD trajectory calculations for 30 ps with occasional temperature scaling, which was followed by a MD run of 100 ps with no temperature scaling.

The intensity of the IR spectrum was calculated by[54]

$$I(\omega) = \frac{\omega}{3\hbar\varepsilon_0 Vc} \tanh\left(\frac{\beta\hbar\omega}{2}\right) \int_{-\infty}^{\infty} dt\, e^{-i\omega t}\langle \mathbf{M}(t)\cdot\mathbf{M}(0)\rangle \tag{21}$$

where V is the volume of system, c is the speed of light, ε_0 is the vacuum dielectric permittivity, and $\beta = 1/k_BT$ with k_B and T being the Boltzmann constant and temperature,
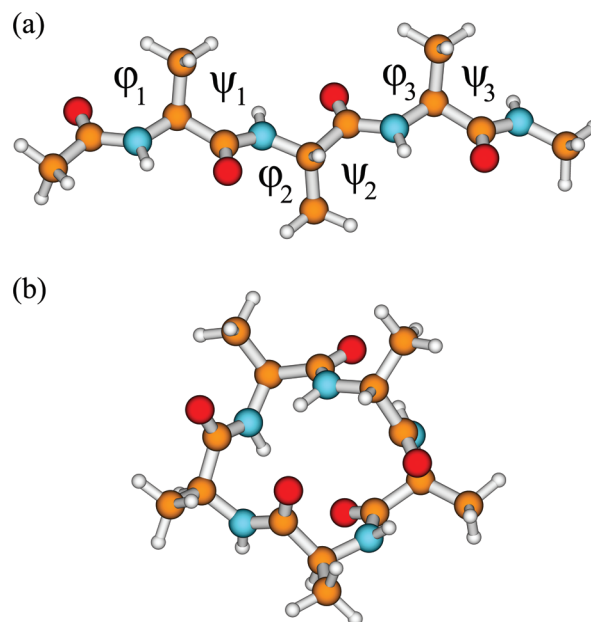
(a)

(b)

Figure 4. Starting conformations of tetrapeptide (a) and pentapeptide (b) in MD simulations. For tetrapeptide, the numbering of the dihedral angles is shown.

respectively. $\langle \mathbf{M}(t)\cdot\mathbf{M}(0)\rangle$ is the time correlation function (TCF) of the system dipole moment $\mathbf{M}$. Since we focused on the effect of charge polarization induced by the intra- and intermolecular interactions on the IR spectra, the contributions of the dipole moment of water solvent were subtracted from the system dipole moment in calculating the IR spectra.

To analyze the polarization effect on the spectrum, the dipole moment of the peptide in solution is decomposed into the gas-phase dipole moment $\mathbf{M}_G^U$ and the solvent-induced one $\mathbf{M}_I^U$ as

$$\mathbf{M}^U(t) = \mathbf{M}_G^U(t) + \mathbf{M}_I^U(t) \tag{22}$$

where the superscript U means the peptide dipole moment.[33] Note that the gas-phase dipole moment includes the effect of induced charge due to the intramolecular electrostatic interaction. The TCF is thus written as

$$C^U(t) = \langle \mathbf{M}^U(t)\cdot\mathbf{M}^U(0)\rangle = C_{GG}^U(t) + C_{II}^U(t) + C_{GI}^U(t) \tag{23}$$

where the first and second terms are the contributions from the gas phase and solvent-induced dipoles, respectively, and the third one is the cross term between them.

3.2. Results and Discussion. *Dihedral Angle Distribution.* Figure 5 shows the normalized distributions for the backbone dihedral angles (ϕ, ψ) of tetra- and pentapeptides, respectively. The dihedral angles for the main chain of tetrapeptide is numbered from the N terminus, as shown in Figure 4. For the tetrapeptide, the dihedral angle distributions show a large difference between the gas and solution phases despite the initial structures being the same. This is because the isolated tetrapeptide favors a compact form stabilized by the intramolecular hydrogen bonds, while the solvated peptide retains the initial extended form, which is stabilized
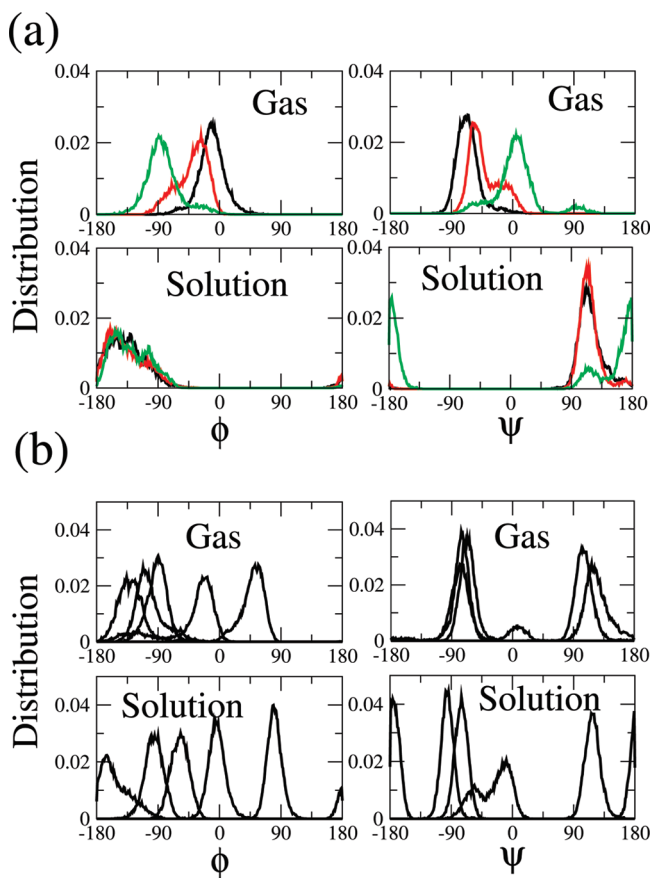
(a)

(b)

Figure 5. (a) Normalized distributions of backbone dihedral angles for tetrapeptide. Distributions of dihedral angles ϕ_1, ϕ_2, and ϕ_3 are written with black, red, and green, respectively. The right panel is the distributions of dihedral angles, ψ_1, ψ_2 and ψ_3. (b) Normalized distributions of dihedral angles for cyclic pentapeptide. The dihedral angles $\phi_1-\phi_5$ and $\psi_1-\psi_5$ are not distinguished.

by the hydrogen bondings with the surrounding water molecules. For the pentapeptide, there was no drastic conformational change within the present simulation time, 100 ps. In addition, a significant difference in distribution of the dihedral angles between the gas and solution is not observed, indicating that the dihedral angles of cyclic pentapeptide are rather restricted by taking the cyclic form.

Charge Fluctuation and Dipole Moment. The atomic site charges in the peptide fluctuate due to the thermal fluctuation of the electrostatic environment. We calculated the average site charges and their fluctuations for the four backbone atomic sites, N, H, C, and O, and the α- and β carbons, C_α and C_β. Similar analyses were performed by Liu et al.[56] for crambin with the QM/MM method, and for several small proteins by Patel and Brooks[17] with the fluctuating charge model. They showed that the polarization of atomic site charges is characterized by the chemical environments around the atoms due to the protein sequence. Since the alanine polypeptides were chosen as the solute in the present calculations, the effects of intra- and intermolecular hydrogen bonding on the site charges and their fluctuations can be discussed because the chemical environments around each backbone atom are very similar between the units. Here, we focused on the cyclic pentapeptide because the environment

around all of the alanine units is equivalent by the symmetry. Figure 6 shows the average charges at the backbone atomic sites in each unit and their fluctuations obtained from the 100 ps trajectory records, where the left and right panels are those in the gas phase and in solution, respectively. It is noted that the site charge in each unit becomes the same if we take a large number of trajectories adequate to represent the thermal average. However, since we took one trajectory starting from the particular initial geometry and there was no conformational change during the 100 ps trajectory period, the site charge depends on the electrostatic field around the atom and has different values for different units, as seen in Figure 6. The average charge shows weak dependency on the amino acid unit in the gas phase, indicating that there is no strong correlation between the atomic charges and dihedral angle distributions. On the other hand, the difference in average charge becomes appreciable in the solution. For example, the site charge on N in unit 4 is larger than that in unit 1 by 0.5, implying that the solvated condition is different between the two atoms. As shown in Figure 6b, the charge fluctuation in solution is larger than that in the gas phase for all of the sites. It is also noted that the fluctuation at the β carbon C_β is approximately 1/10 of that at the α carbon C_α, because the large diagonal CRK element of C_β is canceled by the three equivalent large $C_\beta-H$ off-diagonal elements under a nearly uniform electrostatic environment around the methyl group.

Figure 7 shows the distributions of the dipole moments of tetra- and pentapeptides in solution. The dipole moments at the peak position are almost the same, 11.4 D, for both the peptides. In order to see the role of charge polarization on the dipole moments, we calculated the dipole moments in the gas phase using the CRK charges as well as the reference ones. Note that, for calculating the dipole moments, we used the same trajectory records generated by the MD simulations in solution with the present polarizable force field. As seen in Figure 7, the dipole moment at the peak position increases by 3.3 D due to the intramolecular electrostatic interaction for the tetrapeptide. The solvent effect further enhanced the dipole moment by 1.7 D. For the cyclic pentapeptide, the increase of the dipole moment by the intramolecular interaction, 0.6 D, is smaller than that of the tetrapeptide. On the other hand, the solvent-induced contribution to the dipole moment is larger for pentapeptide, 3.1 D, than that of tetrapeptide. Compared with the pentapeptide, the distribution of the dipole moment of the tetrapeptide is broader. This is because the geometry of the tetrapeptide is flexible even in a solvent, while that of the pentapeptide is restricted to retain the cyclic form.

We further carried out ab initio electronic structure calculations for the tetra- and pentapeptides both in the gas and solution phases, and the resultant dipole moments were compared with those from the present CRK based model. For the calculations in solution, we employed the reference interaction site model self-consistent field (RISM-SCF) method, where the HF wave function with the 6-31G** basis set was used to describe the solute electronic structures, and the RISM integral equation was solved with the standard model of the water molecule.[55] From the MD trajectory
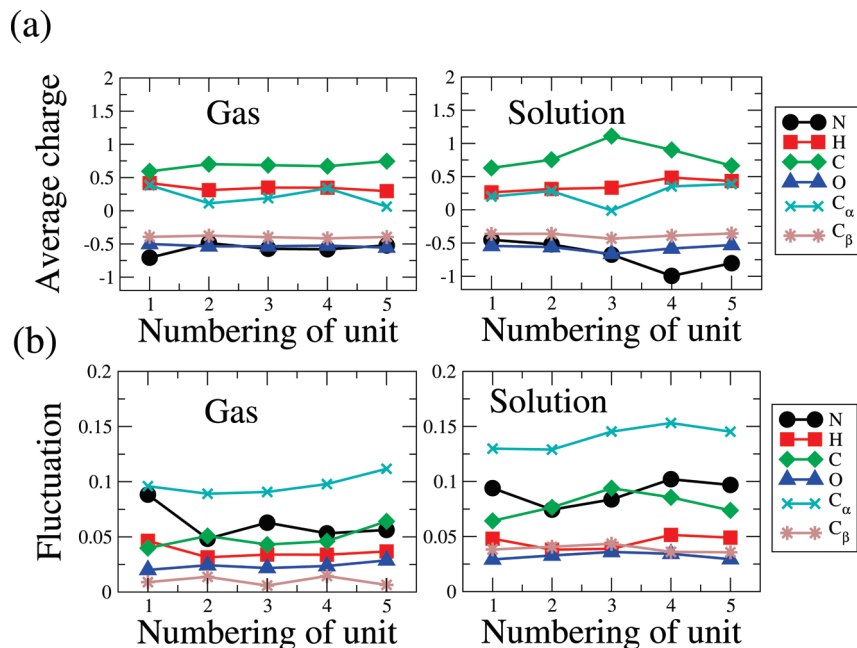
(a)

(b)

Figure 6. Average charges (a) and fluctuations (b) for backbone sites, N, H, C, and O, and α and β carbons, C_α and C_β, for cyclic pentapeptide in the gas phase (left) and solution (right). The *x* axis represents the numbering of the amino acid unit.
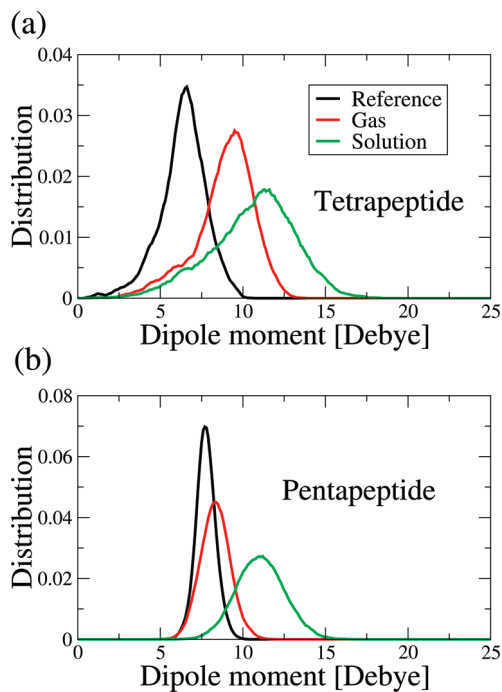
(a)

(b)

Figure 7. Normalized distributions of dipole moment for tetrapeptide (a) and pentapeptide (b) in solution. Black, red, and green correspond to reference, intramolecular induced, and intramolecular induced plus intermolecular induced dipole distributions, respectively.

records, we picked up one geometry for each peptide whose dipole moment is close to that of the peak position in Figure 7. The ab initio dipole moments for the tetrapeptide were 8.6 and 9.2 D in the gas and solution phases, respectively. The corresponding CRK values are 9.7 and 11.4 D, respectively. For the cyclic pentapeptide, the ab initio dipole moments were 9.3 and 11.3 D in the gas phase and solution, respectively, which are close to the CRK values of 8.3 and 11.4 D. These results indicate that the present model provides

a reasonable estimate of charge polarization induced by the intra- and intermolecular hydrogen bondings. It is noteworthy that the dipole moments using the original AMBER99 charge set are 6.4 and 7.7 D for the tetra- and pentapeptides, respectively.

Solvation Structures. In order to examine the effect of peptide−water solvent hydrogen bondings on the charge distribution, we examined the solvation structures. We first calculated the radial distribution functions (RDFs) between the N, H, C, and O atoms in peptides and the water O atoms and then estimated the radius of the sphere determining the first solvation shell for 24 (tetra) and 30 (penta) atoms. The radii thus obtained are 4.5, 3.0, 4.0, 5.0, 4.0, and 5.0 Å for O, H, N, C, H(−C_α), and methyl C atoms, respectively. The corresponding values for pentapeptide are the same as those for the tetrapeptide. Next, we calculated the number of H_2O molecules by counting the numbers of H_2O in the first solvation spheres by checking a double counting of H_2O which is in more than two solvation spheres. The calculated number of H_2O in the first solvation shell was estimated to be 49−62 and 48−61 for the tetra- and peptapeptide, respectively. Since the cyclic pentapeptide has intramolecular hydrogen bondings, the number of coordinated H_2O becomes almost the same as that of the extended tetrapeptide in spite of the difference of the number of amino acid units between the two peptides.

Figure 8 compares the distribution of H_2O dipole moments in the first solvation shell of pentapeptide with that of the bulk water obtained by H_2O molecules far from the peptide atomic sites with a distance 10−11 Å. As seen in the figure, the average dipole moment of H_2O in the bulk water is 2.4 D, which is larger than the gas phase one, 1.9 D, obtained from the reference charges (see section 3.1). It is found that the dipole moment of H_2O around the N atoms is largely enhanced from the bulk one. On the other hand, the distributions of dipole moment around the other peptide
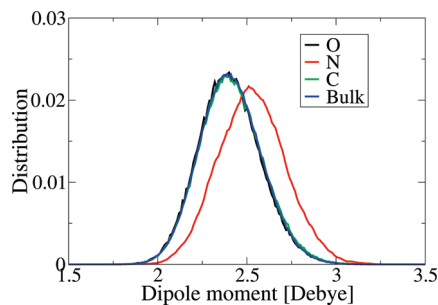
Figure 8. Distributions of water dipole moment in the first solvation shell of carbonyl oxygen (black), amide nitrogen (red), and carbon in the methyl group (green). The distribution in bulk water is also shown as a reference (blue).

atoms are very close to that of the bulk water. It is noted that the dipole moment distribution around the tetrapeptide was similar to that of the bulk solvent, though the dipole moment of the H_2O molecule forming a hydrogen bond with the NH or CO group is largely enhanced.

We further obtained the RDFs between the H (in NH)−O, O (in CO)−H, and C (in CH_3)−O site pairs to see the hydrogen bondings and hydrophobic interactions between the peptides and water. The dependence on the amino acid units was weak for the tetrapeptide because the extended form is favored in aqueous solution. For the cyclic pentapeptide, we observed a strong unit dependence of the number of hydrogen bondings, because the formation of the hydrogen bonding between the carbonyl O and water H atoms is prevented by the intramolecular hydrogen bonding. For the RDFs between the H atom in NH bond and water O atom in Figure 9b, we found two higher peaks in the first solvation shell. These peaks represent the strong hydrogen bonding in units 4 and 5, which is correlated to the enhancement of negative charges on the N atoms in units 4 and 5 in Figure 6. The sharp peaks observed in Figure 9c indicate the hydrohobic interaction of the methyl group with the water solvent.

IR Spectrum. Figure 10 shows the IR spectra up to 2000 cm^{-1} for the tetra- and pentapeptides, where the left and right panels are those in the gas and solution phases, respectively. The peak positions are almost the same between the two peptides in the gas phase. The two split bands at 1670 and 1770 cm^{-1} are attributable to the amide I mode, which arises mainly from the C=O stretching vibrations. Such a band splitting is not observed for the spectrum of N-methylacetamide (NMA) in the gas phase[43,45] because it is attributed to the coupling of the C=O stretching motions between the different units.[46,47,57,58] The split band at ~1550 cm^{-1} is attributable to the amide II mode, which is mainly described by the out-of-phase combination of the N−H in-plane bend and C−N stretching vibrations. The amide III band emerges in the region 1200−1400 cm^{-1} and is rather complex, as seen in Figure 10. Compared with the amide I band, the intensities of amide II and III bands are too small. This is because the dependence of atomic charges on the nuclear coordinates[29] is not included in the present model and thus the change of dipole moment due to the stretching of the amide bond is largely underestimated.[43] The most prominent band around 790 cm^{-1} is the amide V band, which comes
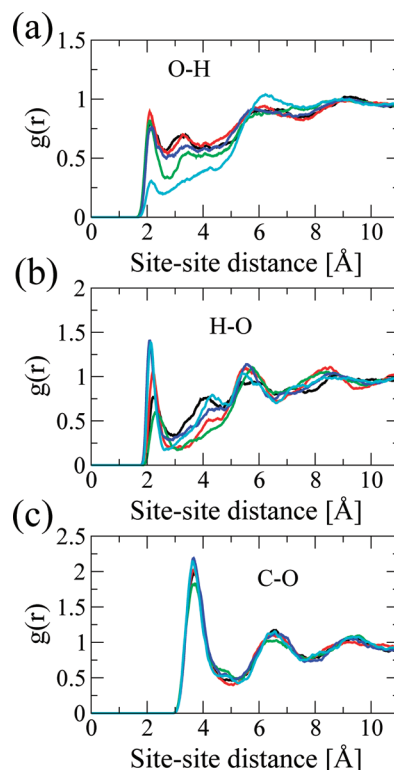

Figure 9. Radial distribution functions between (a) carbonyl oxygen and water H, (b) amide hydrogen and water O, and (c) methyl carbon and water O, respectively. The black, red, green, blue, and cyan curves correspond to the amide acid units from 1 to 5, respectively.

mainly from the out-plane N−H bending motion. The amide V band of pentapeptide shows splitting, indicating that the band profile is sensitive to the intra- and intermolecular environment. The neighbor band at 680 cm^{-1} is assigned to the amide IV and VI modes, which are composed of the out-of-plane C=O bend and the NCO deformation, respectively.

Compared with the gas phase, the shape of each band in aqueous solution becomes broader, and the intensity is largely enhanced. It is notable that the height of the amide V band of tetrapeptide does not increase by solvation, while that of pentapeptide becomes about twice that in the gas phase. In order to clarify the origin of such a difference in the amide V band shape between the tetra- and pentapeptide due to solvation, we decomposed the spectra into the three contributions, that is, the gas phase and solvent-induced self-terms and their cross term, eq 23. The results are displayed in the left panels of Figure 11, where we can see that the contribution of the solvent-induced self-term is small for the terapeptide, though such a term has the largest contribution for the pentapeptide. We further analyzed the vibrational density of state (VDOS) for the N−H bending motion for both the peptides by representing the correlation function of the N−H bond vector as the sum of two terms:

$$\langle \mathbf{R}(t) \cdot \mathbf{R}(0) \rangle / \langle \mathbf{R}(0) \cdot \mathbf{R}(0) \rangle = \sum_i \langle \mathbf{r}_i(t) \cdot \mathbf{r}_i(0) \rangle / \langle \mathbf{R}(0) \cdot \mathbf{R}(0) \rangle$$

$$+ \sum_i \sum_{j(\neq i)} \langle \mathbf{r}_i(t) \cdot \mathbf{r}_j(0) \rangle / \langle \mathbf{R}(0) \cdot \mathbf{R}(0) \rangle$$
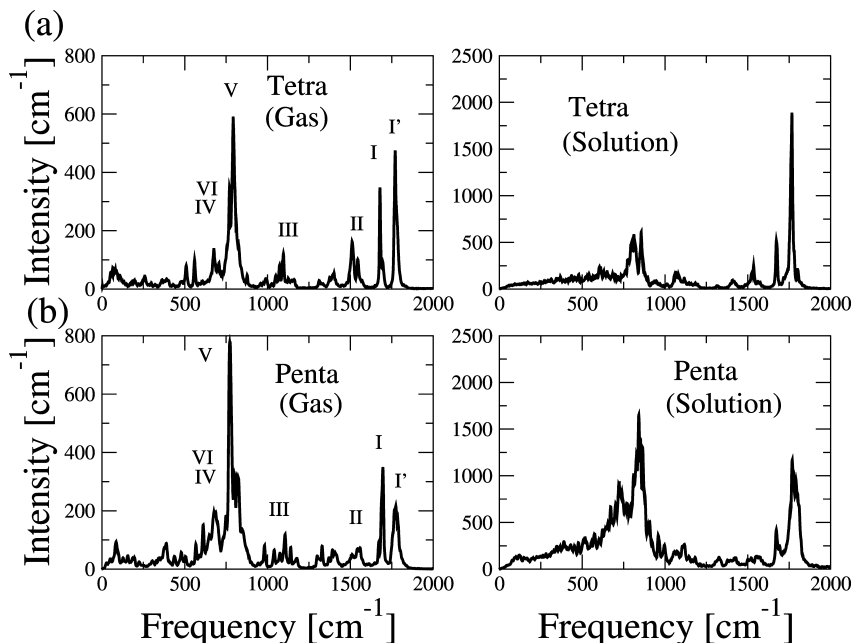
$$(24)$$

Figure 10. IR spectra for (a) tetrapeptide and (b) cyclic pentapeptide. Left and right panels are in the gas phase and in solution, respectively.
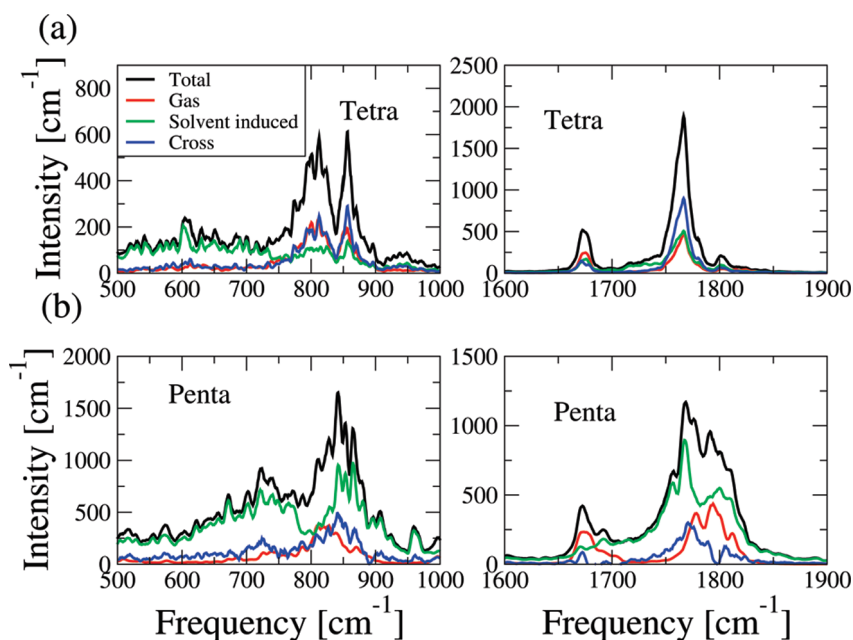

Figure 11. IR spectra of (a) tetra and (b) cyclic pentapeptide in solution in the regions of 500−1000 cm^{-1} (left) and 1600−1900 cm^{-1} (right). Decomposed into gas- and solvent-induced self-terms and their cross term.

where r_i is the N−H bond vector in the ith amino acid unit and R (= $\sum_i r_i$) is the sum of N−H bond vectors. Thus, the VDOS spectrum is decomposed into the diagonal and off-diagonal contributions. The calculated VDOSs of tetra- and cyclic pentapeptides in the region 500−1000 cm^{-1} are shown in Figures S2 and S3 in the Supporting Information. Although both the diagonal and off-diagonal terms have positive intensities for the pentapeptide, the off-diagonal term has a negative contribution for the tetrapeptide, indicating that the enhancement of the amide V band for tetrapeptide is suppressed by the cancellation between the diagonal and off-diagonal terms.

The amide I band is known to be strongly affected by the dihedral structure of the backbone, and its frequency shift is utilized to probe the secondary structures of protein. The aqueous solvation effect on the band shape has been also discussed. As shown in Figure 10, the intensity is largely enhanced due to the solvation both for the tetra- and pentapeptides, and the increase of band intensity at 1770 cm^{-1} is notable. The right panels of Figure 11 show the components of IR spectra in the region 1600−1900 cm^{-1}, where we can see that the largest contribution to the intensity enhancement for the tetrapeptide is the cross term between the gas phase and solvent-induced ones. On the other hand,

the solvent-induced term is dominant, and the cross term is minor for the pentapeptide. One of the possible reasons for the small cross term of pentapeptide compared with the other two terms is that the directions of dipoles coming from the intramolecular and solvent-induced terms compensate each other. It is also noted that the band around 1770 cm^{-1} of pentapeptide shows a red shift from the gas-phase band due to the contribution from the solvent-induced self-term.

4. Conclusions

In the present paper, we developed a polarizable MM force field for polypeptides and proteins employing the CRK model. The effective charges and CRK matrices of polypeptides were constructed from those determined for the constituent amino acids. The electrostatic interaction potential thus obtained was combined with the AMBER99 force field. Furthermore, in order to reproduce the ab initio optimized geometries and relative energies of several conformers for 20 neutral and five charged dipeptides, we refitted the parameters of torsional potentials.

To evaluate the present MM force field, we applied it to MD simulations of the extended alanine tetrapeptide and cyclic pentapeptide. For the pentapeptide, the average site charges on the atoms in the main chain and their fluctuations in aqueous solution were compared with those in the gas phase in order to see the effect of intra- and intermolecular hydrogen bonding on the charge polarization. We also compared the dipole moments of the polypeptides by the present CRK-based model with those by ab initio calculations. As a result, the present model was shown to provide reasonable charge distributions both in the gas phase and in solution.

Finally, we calculated the IR spectra of alanine tetra- and pentapeptides in the gas phase and in solution in order to see the effect of intra- and intermolecular hydrogen bonding on the IR spectral profiles. For both of the peptides, split amide I modes were observed, which are attributed to the coupling of CO stretching modes between the different amino acid units. We also found that the intensity of the amide V mode for the tertapeptide does not increase due to the solvation, though that of pentapeptide is largely enhanced in solution.

It is noted that the present model should be regarded as the first step for the development of a polarizable protein force field employing the CRK method. In order to improve the model, we need to apply it to many problems such as solvation structure and free energy calculations and refine the parametrizations.

Supporting Information Available: Figure S1 shows the schematics of structures for amino acids and terminal species considered in this work. The reference site charges for 20 neutral and five charged amino acids as well as terminal species are given in Tables S1–S27. The CRK matrix elements are also listed in Tables S1–S27. The fitted parameters for the torsional potentials are summarized in Table S28. The ab initio optimized values of key dihedral angles for stable conformers of dipeptides are summarized in Tables S29–S46. The rms deviations of relative energies and dihedral angles from the ab initio reference values are given in Tables S47–S63. The decomposition analyses of the VDOS spectra into the contributions from the diagonal and off-diagonal terms are shown in Figures S2 and S3. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(2) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(3) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(4) Yin, D.; Mackerell, A. D. *J. Comput. Chem.* **1998**, *19*, 334–348.

(5) Feller, S. E.; Mackerell, A. D. *J. Phys. Chem. B* **2000**, *104*, 7510–7515.

(6) Klauda, J. B.; Brooks, B. R.; Mackerell, A. D.; Venable, R. M.; Pastor, R. W. *J. Phys. Chem. B* **2005**, *109*, 5300–5311.

(7) Daura, X.; Mark, A. E.; Gunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535–547.

(8) Oostenbrink, C.; Villa, A.; Mark, A. E.; Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(9) Schuler, L. D.; Daura, X.; Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

(10) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.

(11) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. *J. Phys. Chem.* **1994**, *98*, 13077–13082.

(12) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(13) Dang, L. X.; Rice, J. E.; Caldwell, J.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 2481–2486.

(14) Cieplak, P.; Caldwell, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1048–1057.

(15) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515–1531.

(16) Rick, S. W.; Stuart, S. J.; Bern, B. J. *J. Chem. Phys.* **1994**, *101*, 6141–6156.

(17) Patel, S.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1–15.

(18) Patel, S.; Mackerell, A. D.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504–1514.

(19) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

(20) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025–3039.

(21) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2005**, *1*, 153–168.

(22) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(23) Harder, E.; Kim, B.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2005**, *1*, 169–180.

(24) Kim, B.; Young, T.; Hardrer, E.; Friesner, R. A.; Berne, B. J. *J. Phys. Chem. B* **2005**, *109*, 16529–16538.

(25) Soto, P.; Mark, A. E. *J. Phys. Chem. B* **2002**, *106*, 12830–12833.

(26) Jiao, D.; Zhang, J.; Duke, R. E.; Li, G.; Schnieders, M. J.; Ren, R. *J. Comput. Chem.* **2009**, *30*, 1701–1711.

(27) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *29*, 6290–6295.

(28) Morita, A.; Kato, S. *J. Am. Chem. Soc.* **1997**, *119*, 4021–4032.

(29) Lu, Z.; Yang, W. *J. Chem. Phys.* **2004**, *121*, 89–100.

(30) Ishida, T.; Morita, A. *J. Chem. Phys.* **2006**, *125*, 074112.

(31) Isegawa, M.; Kato, S. *J. Chem. Phys.* **2007**, *127*, 244502.

(32) Morita, A.; Kato, S. *J. Chem. Phys.* **1998**, *108*, 6809–6818.

(33) Iuchi, S.; Morita, A.; Kato, S. *J. Phys. Chem. B* **2002**, *106*, 3466–3476.

(34) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(35) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781–790.

(36) Blatt, H. D.; Smith, P. E.; Pettitt, B. M. *J. Phys. Chem. B* **1997**, *101*, 7628–7634.

(37) Young, W. S.; Brooks, C. L., III. *J. Mol. Biol.* **1996**, *259*, 560–572.

(38) Takano, M.; Yamato, T.; Higo, J.; Suyama, A.; Nagayama, K. *J. Am. Chem. Soc.* **1999**, *121*, 605–612.

(39) Hu, H.; Elstner, M.; Hermans, J. *Proteins* **2003**, *50*, 451–463.

(40) Cui, Q.; Smith, H. *J. Chem. Phys.* **2002**, *118*, 279–290.

(41) Wang, Z.-X.; Duan, Y. *J. Comput. Chem.* **2004**, *25*, 1699–1716.

(42) Gao, J.; Freindorf, M. *J. Phys. Chem. A* **1997**, *101*, 3182–3188.

(43) Gaigeot, M. P.; Vuilleumier, R.; Sprik, M.; Borgis, D. *J. Chem. Theory Comput.* **2005**, *1*, 772–789.

(44) Cho, M.; Yang, S. *J. Chem. Phys.* **2005**, *123*, 134503.

(45) Bouør, P. *J. Chem. Phys.* **2004**, *121*, 7545–7648.

(46) Moran, A. M.; Park, S.-M.; Mukamel, S. *J. Chem. Phys.* **2003**, *118*, 9971–9980.

(47) Bouř, P.; Keiderling, T. A. *J. Chem. Phys.* **2003**, *119*, 11253–11262.

(48) Torii, H. *J. Phys. Chem. B* **2007**, *111*, 5434–5444.

(49) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341–350.

(50) Ewing, C. S.; Walman, M.; Maple, J. R. *J. Phys. Chem. A* **2002**, *106*, 326–334.

(51) Boehm, H. J.; Brode, S. *J. Am. Chem. Soc.* **1991**, *113*, 7129–7135.

(52) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908–5920.

(53) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F.; Hermans, J. *Intermolecular Force*; Reidel: Dordrecht, The Netherlands, 1981; pp 331−342.

(54) McQuarrie, D. A. *Statistical Mechanics*; Harper Collins Publishers: New York, 1976; pp 60−65.

(55) Ten-no, S.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1994**, *100*, 7443–7453.

(56) Liu, H.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Hermans, J.; Yang, W. *Proteins* **2001**, *44*, 484–489.

(57) Ganim, Z.; Chung, H. S.; Smith, A. W.; DeFlores, L. P.; Jones, K. C.; Tokmakoff, A. *Acc. Chem. Res.* **2008**, *31*, 432–441.

(58) Barth, A.; Zscherp, C. *Q. Rev. Biophys.* **2002**, *35*, 369–430.

JCTC Journal of Chemical Theory and Computation

A Consistent Force Field for the Carboxylate Group

Maxim Tafipolsky and Rochus Schmid*

*Lehrstuhl für Anorganische Chemie 2, Organometallics and Materials Chemistry,
Ruhr-Universität Bochum, Universitätsstr. 150, D-44780 Bochum, Germany*

Abstract: In the light of the important role played by the carboxylate group in bio- and coordination chemistry, its consistent and reliable parametrization for molecular simulations is crucial. The experimental vibrational spectra of three carboxylate anions (formate, acetate, and benzoate) both in the gas phase and in the condensed phase (as sodium salts) are interpreted on the basis of high-quality ab initio calculations. The interaction with the counterion (metal cation) is shown to be of major importance in the interpretation of the spectral features of the carboxylate group both in the solid state and in aqueous solution. Previous attempts to parametrize the carboxylate group within the molecular mechanics approach is critically reviewed, and a new set of the consistent valence force field parameters based on first principles calculations is proposed, which is able to reproduce accurately both the structure and the dynamics of the carboxylate moiety both free and coordinated with metal cations.

1. Introduction

Despite the fact that the carboxylate group is an important functionality abundant in biological systems, being one of the major intracellular anions, its parametrization within the molecular mechanics (MM) approach has been rarely done consistently. If both the structure and dynamics of the carboxylate group, either free (isolated) or chelated with metal ions, are required, then a systematic derivation of the force field (FF) parameters is needed. In the course of the development of a FF for the relatively new class of nanoporous metal−organic frameworks (MOFs),[1] we have faced the problem of accurate parametrization of the metal−carboxylate interactions, which are crucial for understanding the structural, thermal, and elastic properties of these materials. Unfortunately, the parameters available in the literature are not able to deliver reliable dynamic quantities (e.g., vibrational frequencies). Hence, there is a need for a systematic strategy for the parametrization of this important organic group.

Interestingly, the most simple carboxylate anion, formate, was used in the late 1980s as an example to demonstrate the usefulness of using first principles calculations for the derivation of FF parameters either alone[2] or in combination with experimental data.[3] Since then, a number of attempts

have been aimed at the development of the parameters for the carboxylate group using various functional forms. Utilizing a "building block" approach, where any large system is broken down structurally into key functional components, which were parametrized separately, Kirschner et al.[4] have studied the conformational flexibility of amino acid zwitterions on the basis of their parameters for the carboxylate group derived within the MM3 force field. To develop force constants, they used Møller−Plesset second-order (MP2) perturbation calculations on small carboxylates. Whereas structural data were well reproduced, their inability to match vibrational frequencies of the carboxylate group (mainly symmetric and asymmetric stretchings), as will be shown below, was due to the absence of some important interactions (cross-terms) between internal coordinates in the MM3 energy expression, such as the stretch−stretch term. We should note, however, that the importance of these cross-terms has been recognized, and they were implemented in the MM4 force field[5] for the conjugated systems like benzene and butadiene. Moreover, Lii[6] in his study based on the MM4 force field has shown that including the C−O/C−O stretch−stretch cross-term improves significantly the agreement with the high-level ab initio calculated frequencies (B3LYP/6-31G(d, p)) for carboxylic acids. Below, we will show that this particular cross-term plays a crucial role in

---

correctly reproducing the C−O stretchings in the carboxylate group as well.

Making use of first principles calculations as the basis for the FF development, one has to be sure that a chosen level of theory is reliable; that is, the results produced are in agreement with high-quality experimental data within specified uncertainties. Recently, thanks to experimental advances, accurate experimental vibrational frequencies for the free carboxylate anions became available. Forney et al.[7] investigated by IR spectroscopy some important fundamental vibrations of the formate anion and its deuterated analog trapped in a solid neon matrix at 5 K, whereas Oomens and Steill, using infrared multiple photon dissociation action spectroscopy, observed, for the first time, the stretching modes of the free benzoate[8] and acetate[9] anions in the gas phase.

The carboxylate moiety is known to have a very diagnostic infrared fingerprint: the symmetric and asymmetric C−O stretch modes are extremely sensitive to the environment.[10] These modes, and in particular their separation, have therefore been widely used in condensed-phase spectroscopic studies to determine the coordination mode in carboxylate salts and solutions.[10−12] Since some parametrizations were based on the experimental data (structure and frequencies) measured for the condensed phase, the bias due to the environment should be accounted for (counterions or solvent).

In this work, to derive the needed FF parameters, we use the results of first principles calculations, the reliability of which is judged from a comparison with experimental data and other high-level calculations where available.

2. Force Field Derivation

Our model systems include three anions, formate, acetate, and benzoate, and their chelated complexes with alkali (Li and Na) and transition metal (Cu and Zn) cations. Their geometries were fully optimized by a number of methods ranging from density functional theory (DFT), Møller−Plesset second-order (MP2) perturbation theory,[13] and coupled-cluster (CCSD[14] and CCSD(T)[15]) theory. As DFT, we use the hybrid functional, B3LYP.[16−18] Only valence electrons were correlated in the coupled-cluster calculations. The augmented correlation consistent basis sets, aug-cc-pVXZ (X = D, T, Q),[19] were employed for all atoms if not stated otherwise. For the Zn and Cu atoms, energy-consistent pseudopotentials denoted as (aug-)cc-pVXZ-PP (X = D, T),[20−22] obtained from the William R. Wiley Environmental Molecular Sciences Laboratory Basis Set Exchange,[23,24] were employed. In addition to the calculations in a vacuum, the conductor polarizable continuum model (CPCM) was used to take the effects of a dielectric medium (water) into account, with the cavity represented by the United Atom Topological Model applied on radii optimized for the PBE0/6-31G(d) level of theory (RADII = UAKS). All calculations were carried out with the Gaussian program package.[25] The optimized structures were confirmed to be true minima by vibrational frequency calculations. Anharmonic frequencies were calculated by numerical differentiation along normal modes as implemented in the Gaussian program package.[26]

At the optimized geometries, the Hessian matrices in Cartesian coordinates were calculated using analytical second derivatives and then projected into a redundant set of internal coordinates by means of the Moore−Penrose generalized inverse technique and further used in the derivation of the FF parameters. By transforming the Hessian matrix from Cartesian to internal coordinates, rows of the **B** matrix that correspond to translations and rotations (Eckart coordinates) are formed, and the Hessian matrix is orthogonalized with respect to these coordinates such that both translational and rotational degrees of freedom are completely projected out.

The functional form of the MM3(2000) force field of Allinger et al.[27] was used as implemented in the TINKER program package.[28] We have modified the original MM3 stretch−bend cross-term by introducing two different force constants, k_{sb}^a and k_{sb}^b, for two bonds that comprise the same bond angle as suggested by Mapple et al.[29] and implemented a new stretch−stretch cross-term in the form used previously.[5,6,29] Their functional forms are given below, where r_{ref} and α_{ref} are the reference values for the bond length and bond angle, respectively:

For stretch−stretch cross-terms:

$$E_{ss} = k_{ss}(r^a - r_{ref}^a)(r^b - r_{ref}^b) \tag{1}$$

and for stretch−bend cross-terms:

$$E_{sb} = k_{sb}^a(r^a - r_{ref}^a)(\alpha - \alpha_{ref}) + k_{sb}^b(r^b - r_{ref}^b)(\alpha - \alpha_{ref}) \tag{2}$$

We concentrate on the FF parameters pertaining to the carboxylate group, namely, C−O stretchings, O−C−O in-plane and out-of-plane bendings, and (for the benzoate anion) twisting around the C_{carb}−C_{ph} bond. Our strategy is to project both structural and curvature (Hessian matrix) data into the redundant internal coordinate system. From a practical point of view, the use of redundant coordinates is advantageous, since the point group symmetry of the carboxylate moiety and corresponding force field is preserved; that is, all symmetrically equivalent bonds or angles in the group have the same force constants. We avoid any symmetrization (producing a nonredundant set of internal coordinates) or "localization"[30] steps (resulting in maximally diagonal force constants in dependent angle-bending coordinates), which is completely in line with the general MM philosophy where redundant coordinates are used. The problem with the angular redundancies does not occur[31] because all geometry optimizations and Hessian calculations are carried out in Cartesian coordinates, and we use exactly the same (complete) set of internal coordinates in the transformation of the Hessian. In addition, due to the redundancy relationship among the three valence angles around the trigonal-planar C_{carb} atom, the values for some cross-terms (see Figure 1) are reproduced automatically and there is no need to parametrize them explicitly. These are all bend−bend and CO/RCO (and CR/OCO) stretch−bend interactions (R = H, C).

3. Results and Discussion

First, we present and discuss the structural data for the simplest carboxylate, the formate anion, which was the
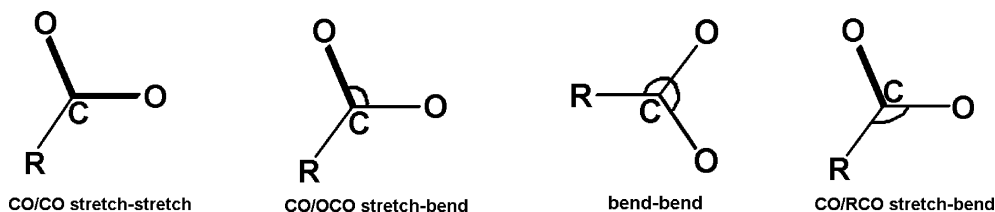
Figure 1. Some important cross-terms in the carboxylate group.

Table 1. Selected Structural Data for the Formate Anion

method	r(C–O), Å	∠O–C–O, deg
B3LYP/cc-pVDZ	1.2519	131.28
B3LYP/aug-cc-pVDZ	1.2578	130.26
B3LYP/aug-cc-pVTZ	1.2512	130.39
B3LYP/aug-cc-pVQZ	1.2495	130.39
MP2/aug-cc-pVDZ	1.2694	130.15
MP2(full)/aug-cc-pVDZ	1.2685	130.15
MP2/aug-cc-pVTZ	1.2583	130.20
MP2/aug-cc-pVQZ	1.2551	130.15
CCSD/aug-cc-pVDZ	1.2626	130.25
CCSD(T)/aug-cc-pVDZ	1.2689	130.28
MP4/6-311++G(d, p)[a]	1.264	130.48
CCSD(T)/aug-cc-pVTZ[b]	1.258	130.2
CCSD(T)/aug-cc-pVQZ[c]	1.2535	130.18

[a] Reference 32. [b] Reference 33. [c] Reference 34.

Table 2. Selected Structural Data for the Three Carboxylate Anions

method	r(C–O), Å	∠O–C–O, deg
Formate		
B3LYP/aug-cc-pVDZ	1.2578	130.26
B3LYP/aug-cc-pVTZ	1.2512	130.39
MP2/aug-cc-pVDZ	1.2694	130.15
MP2/aug-cc-pVTZ	1.2583	130.20
CCSD(T)/aug-cc-pVDZ	1.2689	130.28
Acetate		
B3LYP/aug-cc-pVDZ	1.2618	128.55
B3LYP/aug-cc-pVTZ	1.2548	128.75
MP2/aug-cc-pVDZ	1.2720	128.64
MP2/aug-cc-pVTZ	1.2609	128.76
CCSD(T)/aug-cc-pVDZ	1.2715	128.73
Benzoate		
B3LYP/aug-cc-pVDZ	1.2591	129.14
B3LYP/aug-cc-pVTZ	1.2524	129.23
MP2/aug-cc-pVDZ	1.2688	129.48

subject of numerous theoretical investigations. The most recent contributions are by Magalhaes et al.,[32] Dixon et al.[33] and Krekeler et al.[34] The most important structural parameters calculated at different levels of theory are compared in Table 1.

Using the results from the highest level of theory (CCSD(T)/aug-cc-pV6Z+aug-cc-pV5Z, all electrons correlated) used in the work of Krekeler et al.[34] as the benchmark, we note that their values (r(C–O) = 1.2501 Å and ∠O–C–O = 130.14°) can be well reproduced using a much cheaper B3LYP/aug-cc-pVTZ level of theory. Augmentation of the basis set with the diffuse functions has been known for quite a long time to be important for reliable predictions of the properties of anions.[35] If we compare the results from our two DFT calculations (see Table 1) with (B3LYP/aug-cc-pVDZ) and without (B3LYP/cc-pVDZ) diffuse functions, the only visible deviation is found for the O–C–O angle (~1°). More drastic changes, however, will be seen below when comparing vibrational frequencies (see also ref 32). The influence of the substituent on the structural parameters of the carboxylate group can be seen from the data presented in Table 2.

One clear trend is evident from the calculated results, namely, the C–O bond distance becomes shorter by ~0.007 Å on going to a larger basis set. At the same time, one can see no pronounced influence of the substituent on the C–O bond distance. We also note that the stretching of the C–O distances results in the closing of the O–C–O angle.

The calculated vibrational frequencies for the three carboxylate anions are compared with the available experimental data in Table 3. We focus here on three normal modes: the symmetric and asymmetric CO stretching and the OCO bending.

By comparing our calculated vibrational frequencies for the formate anion with those of Krekeler et al.,[34] we see

that the B3LYP/aug-cc-pVTZ level of theory is quite adequate for describing dynamics within the carboxylate moiety. Importantly, the harmonic frequencies should be corrected for anharmonicity effects in order to be in agreement with experimental results. This correction is, however, different for various modes. In our case, both symmetric and asymmetric CO stretching modes are appreciably affected by anharmonicity (~30 cm^{-1}), whereas the OCO bending is almost not affected. Overall, very good agreement between our corrected values and the experimental data for the free formate anion gives us confidence in the chosen level of theory (B3LYP/aug-cc-pVTZ). We can reproduce both the absolute values for the CO stretching frequencies and their difference as well (see last column in Table 3). We should note that, whereas the absolute values for the CO stretching frequencies are dependent on the level of theory used, their difference ($\Delta\nu_{as-s}$, see last column in Table 3) is much less prone to systematic errors, such as basis set truncation, anharmonic correction, and so forth. The calculated frequency separation between the CO asymmetric and symmetric stretchings is, therefore, a reliable diagnostic of the chemical environment of a carboxylate group. This separation has been widely used in spectroscopy to elucidate the type of metal ion coordination to the COO moiety.[10] Assuming similar values for the anharmonic correction for carboxylate stretchings, we can see that the agreement between our calculated values and the experimental ones for the acetate and benzoate anions is fairly good. Even a smaller basis set (B3LYP/aug-cc-pVDZ) can deliver reliable results, as shown recently by Oomens and Steill.[8] In their work, they stressed the critical role played by diffuse functions in reproducing stretching modes. In the followup study,[9] all

Table 3. Selected Vibrational Frequencies (in cm^{-1}) for the Three Carboxylate Anions

method	δ(OCO)	ν_{as}(C−O)	ν_{s}(C−O)	$\Delta\nu_{as-s}$
		Formate		
B3LYP/aug-cc-pVTZ	744 (739)[a]	1649 (1621)	1338 (1312)	311 (309)
CCSD(T)/aug-cc-pVQZ[34]	745 (738)	1653 (1619)	1342 (1316)	311 (303)
MP4/6-311++G(d, p)[32]	726	1626	1294	332
exptl[7]	744	1629	1323	306
		Acetate		
B3LYP/aug-cc-pVTZ	863 (838)[b]	1638 (1605)	1347 (1306)[b]	291 (299)
exptl[9]	835	1590	1305	285
		Benzoate		
B3LYP/aug-cc-pVTZ	813[b]	1662	1335[b]	327
exptl[8]	804	1626	1311	315

[a] Anharmonic frequencies are given in parentheses. [b] With the significant contribution from the C$_{carb}$−C stretching.
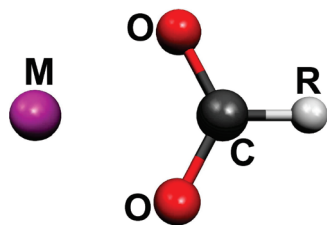
Figure 2. Bidentate complex of metal carboxylate.

calculated harmonic frequencies were scaled by a factor of 0.98, thus correcting for anharmonicity effects. Blue-shifting of the OCO bending mode for both acetate and benzoate anions as compared to the formate anion, as seen from the data presented in Table 3, is mainly due to the strong coupling with the C$_{carb}$−C(R) stretching mode.

Having established a reliable computational framework, it is of importance to go beyond the free anions and to try to simulate the carboxylate moiety in its complexes with metals or in condensed phases where the anion is surrounded by counterions or solvent molecules. Interestingly, various previous attempts to interpret the IR and Raman spectra, measured both in the solid state and in solutions, in terms of the carboxylate fingerprints failed for one or another reason. We argue that taking the counterion (alkali metal cation) into account can explain the main features of the spectra pointing to the minor role played by the solvent. To this end, we have fully optimized the geometry of the carboxylate anions chelated in a simple bidentate fashion by the sodium cation (see Figure 2). The structural data and vibrational frequencies are given in Table 4.

Such a "direct binding"[36] is frequently found in the crystal structures of anhydrous alkali metal salts of formate and acetate (see, for example refs 37−39) or of more complex carboxylates.[40] Recently, Aziz et al.[41] have probed directly the contact ion pair interaction using oxygen K-edge X-ray absorption spectroscopy and molecular dynamics simulations and were able to show that this symmetric bidentate coordination prevails in the aqueous solution of sodium acetate. We note, however, that local environments around a metal cation and a carboxylate group in the crystal structure of sodium acetate trihydrate[42,43] are different from those found in its three anhydrous polymorphs.[37,44] Similar spectral features observed in the solid state of sodium acetate trihydrate and in its aqueous solution can be seen as an

indirect indication of a similar local environment around an acetate moiety in the condensed phase.

Comparing the structural parameters of the free carboxylates (see Table 1) with those in their sodium complexes, we note that the geometry around the C$_{carb}$ atom is perturbed quite significantly by the presence of the counterion. The calculated C−O bond distances are longer than those found in the solid state most probably due to the fact that the Na···O distances in the crystal structures are appreciably larger (2.4−2.7 Å)[37,38] than the calculated ones (2.2 Å), which in turn is a direct consequence of the six-coordinated metal site found in the solid state. In the coordination sphere of the Na cation, two of the six oxygens belong to the same carboxylate (formate or acetate) ion, while the other four oxygens belong to four different ions. Interestingly, the closure of the OCO angle on complexation is well-reproduced by our calculations. From the data presented in Table 4, we confirm the conclusion reached by Keresztury et al.[53] on the importance of taking the counterion into account in order to be able to interpret the vibrational spectra of sodium acetate (see also ref 54). We should also note that the asymmetric stretching mode shows up in the solid state spectra with some scatter (see Table 4). This could be due to the fact that anhydrous sodium acetate is known to exist in three different modifications,[37,44] where a particular atomic arrangement around the acetate anion affects this mode to a larger extent. Taking the solvent (water) implicitly into account within the conductor polarizable continuum model makes the agreement with the experimental frequencies slightly worse. Upon solvation, the most noticeable geometrical change was the lengthening of the Na···O distances (∼0.2 Å), which is, of course, due to the large dielectric constant of water (see also ref 55). These enlarged Na···O separations are still in the range for a "direct binding" (2.3−2.6 Å).[36,40] It is remarkable that this very simple model for the sodium acetate (also used before) can account for the most salient spectral bands. Since sodium acetate has been the subject of numerous spectroscopic investigations, it is informative to compare more closely the calculated harmonic frequencies with their experimental counterparts available in the literature.[49,50]

In solution, in contrast to the crystal structures of anhydrous carboxylate salts, the influence of solvent molecules should be taken into account as well. In aqueous solution, for example, hydration by water molecules com-

Table 4. Selected Structural Data and Vibrational Frequencies (in cm^{-1}) for Sodium Carboxylates

method	r(C–O), Å	∠O–C–O, deg	δ(OCO)	ν_{as}(C–O)	ν_s(C–O)	$\Delta\nu_{as-s}$
			Na-formate			
B3LYP/aug-cc-pVTZ	1.260	125.6	795 (788)[a]	1593 (1560)	1366 (1344)	227 (216)
B3LYP/aug-cc-pVTZ (CPCM)	1.260	125.0	763	1547	1338	209
exptl (solid state)[45]	1.2560(3)[b]	125.50(3)[b]	773	1595	1359	236
exptl (aq. solution)[46]			769[c]	1580	1351	229
			Na-acetate			
B3LYP/aug-cc-pVTZ	1.267	123.6	679	1569	1426	143
B3LYP/aug-cc-pVTZ (CPCM)	1.265	123.4	659	1538	1416	122
exptl (solid state)[47]	1.245, 1.255[d]	123.6[d]	650	1580	1424	160
exptl (solid state)[48]	1.253, 1.257[e]	123.7[e]	658	1562	1424	138
exptl (solid state)[49]	1.250, 1.253[f]	124.3[f]	650	1568	1426	142
exptl (aq. solution)[46]				1550	1416	134
exptl (aq. solution)[49]			653	1556	1420	136
exptl (aq. solution)[50]			654	1551	1416	135
			Na-benzoate			
B3LYP/aug-cc-pVTZ[g]	1.268	123.4	860	1556	1409	147
exptl (solid state)[51]			845	1552	1413	139
exptl (aq. solution)[51]			845	1545	1391	154

[a] Anharmonic frequencies are given in parentheses. [b] Reference 39. [c] Reference 52. [d] Anhydrous sodium acetate.[37] [e] Sodium acetate trihydrate.[42] [f] Sodium acetate trihydrate.[43] [g] Basis set without diffuse functions (cc-pVTZ) were used for all H atoms.
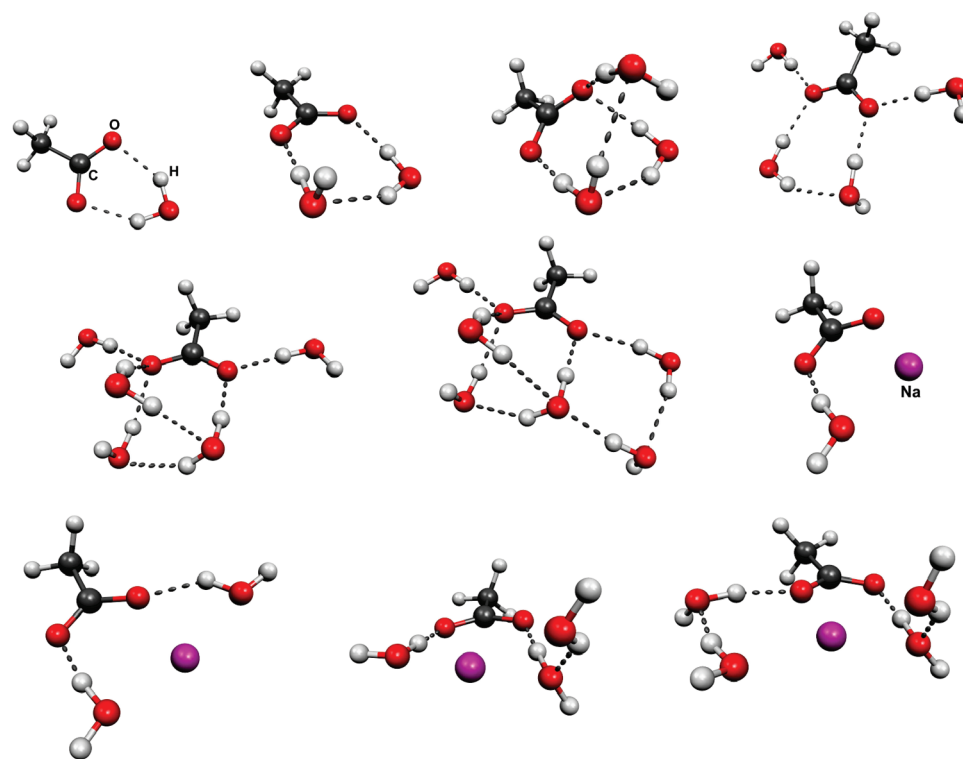

Figure 3. Optimized clusters of acetate anions with the sodium cation and water molecules (B3LYP/aug-cc-pVDZ). Hydrogen bonds are indicated by dash lines.

petes with cation–anion ion pairing. Therefore, it is quite informative to compare the influence of explicit water molecules with that of a counterion in regard to the structural parameters and vibrational frequencies of the carboxylate moiety. To this end, ab initio calculations (B3LYP/aug-cc-pVDZ) on isolated complexes of the acetate anion with one, two, three, four, five, or six water molecules have been carried out. The exploration of the potential energy hypersurface of these clusters was not exhaustive and served mainly to reveal some trends (if any), and only some stable local minima were identified, shown in Figure 3. Recent ab initio calculations and molecular dynamics simulations

indicate that the average number of water molecules in the proximity of the carboxylate groups in formate and acetate (first coordination shell) is in the range of four to six (see, for example, refs 56–58 and references cited therein). Recent time-of-flight neutron diffraction measurements carried out on aqueous sodium acetate solutions indicated that the first hydration shell of an acetate ion comprises four water molecules.[59] Interestingly, our results suggest that an acetate anion can accommodate as many as five water molecules but with a tendency to form hydrogen bonds between water molecules if their number increases further (compare clusters with five and six water molecules in Figure 3). This is

A Consistent Force Field for the Carboxylate Group

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2827**

Table 5. Selected Vibrational Frequencies (in cm^{-1}) for Sodium Acetate

mode[a]	calcd[b]	exptl (Raman, solid state)[49]	exptl (Raman, aq. solution)[49]	exptl (IR, solid state)[47][c]
ν_{as}(C−O)	1569 (1634)	1568	1556	1580
ν_s(C−O)[d]	1426 (1347)	1426	1420	1424
ν(C−C)	929 (863)	935	930	923
δ(OCO)[d]	679 (630)	650	653	650
π(OCO)	622 (599)	617	621	625
ρ(OCO)	465 (436)	478	480	468

[a] The notations ν, δ, ρ, and π are used for the stretching, in-plane bending, rocking, and out-of-plane bending modes, respectively. [b] Harmonic frequencies calculated at the B3LYP/aug-cc-pVTZ level. Values for the free acetate anion are given in parentheses. [c] Anhydrous sodium acetate at 80 K. [d] With a significant contribution from the C−C stretching.

qualitatively in agreement with a recent first-principles (DFT/BLYP) molecular dynamics study of formate anion hydration by Leung and Rempe,[58] where a hydration number of 2.45 per formate oxygen was predicted. Our findings are also in accord with a recent study of da Silva et al.,[60] where explicit solvent molecules were used in the continuum model calculations with the conclusion that five water molecules are enough to represent strong solute−solvent interactions exemplified, among others, by the hydration of ionic solutes such as formate and acetate anions. Table 6 lists the most relevant structural parameters and frequencies for the optimized acetate−water clusters.

Some conclusions can be readily drawn from the data on acetate−water clusters. In terms of both structural parameters and vibrational frequencies, adding water molecules produces a noticeable departure from the values for the isolated anion, which is a direct consequence of the hydrogen bonds being formed between acetate and water molecules. The most evident structural changes are the shortening of the C−C bond distance (and the corresponding blue-shift of the C−C stretching frequency) and the decrease of the O−C−O bond angle, which are in agreement with previous ab initio calculations of Markham et al.[56] The nonequivalence of the two CO bond distances in the clusters with more than one water molecule, due to asymmetric hydrogen bonds, has been pointed out in a recent study by Gojlo et al.[61] (see also refs 12 and 53). The complexation of acetate with one water molecule, the dimer being both experimentally and theoretically well studied (see, for example, refs 62 and 63), does not explain the vibrational features in the carboxylate region observed in an aqueous solution of sodium acetate. We note, however, that, with more water molecules added, an appreciable increase/decrease of the symmetric/asymmetric C−O stretching frequencies becomes evident as compared with the values for the isolated anion. Furthermore, adding one sodium atom to the acetate−water clusters results in the redistribution of the electron charge density within the acetate ion, which manifests itself in a further lowering of the asymmetric C−O stretching frequency and decreasing the frequency separation between the asymmetric and symmetric C−O stretching modes (compare with Table 5). To separate the influence of the cation from that of water molecules to some extent, appropriate crown ethers or cryptands can be used to trap a counterion, thus preventing ion pairing.[64] A direct structural and spectroscopic study of the microsolvation

of the carboxylate anion can be of great help here (see, for example, ref 65). The optimized sodium−acetate−water complexes with one, two, and four water molecules exhibit an asymmetrical arrangement of the sodium atom with respect to the two carboxylate oxygens with the Na···O(carb) distances in the ranges of 2.1−2.3 Å (short contacts) and 3.0−3.6 Å (long contacts). The close contacts between sodium and water molecules in the clusters studied (2.2−2.4 Å) lead to an enhancement of the hydrogen bonding with the carboxylate oxygen atoms thus indirectly influencing the geometry and dynamics of the acetate. This is particularly so in the complex with sodium and three water molecules (see Figure 3), approximately mimicking the local environment around an acetate in the crystal structure of sodium acetate trihydrate,[42,43] where the two calculated Na···O(carb) distances are 3.6 Å and the shortest H−O−H···O(carb) contact is 1.45 Å, whereas in pure acetate−water clusters, our calculations (B3LYP/aug-cc-pVDZ) give H−O−H···O(carb) hydrogen bond distances in the range of 1.6−1.8 Å. This means, that reliable interpretation of the spectra of acetate even in aqueous solution is possible only if the counterion is taken into account. Recently, Park and Woon[66] have studied by DFT (B3LYP/6-31+G(d, p)) the formate anions embedded in water clusters of varying sizes and found that the influence of a counterion (ammonium) should be taken into account to better match the frequencies and intensities of CO stretching modes of the experimental spectra.

Molecular dynamics simulations based on ab initio potentials, which include both the counterions and water molecules explicitly, would be of great help to reliably interpret the structure and dynamics of metal carboxylates in aqueous solution. At present, however, long time simulations utilizing nonempirical potentials are very time-consuming, and therefore, reliable force field parametrizations are of great relevance.

On the basis of the very good overall agreement between our calculated and experimental frequencies, one can conclude that the cation−anion interaction plays an important role both in aqueous solution and in the solid state. We stress, however, that the "free" carboxylate model is incapable of explaining the spectral features observed for metal carboxylates studied here. We note that the inclusion of the effects of solvent implicitly (via CPCM self-consistent field with water as a dielectric medium) perturbed the geometry of the isolated acetate anion quite appreciably; it is, however, not able to reproduce the experimental spectral features (see Table 6). Despite the fact that both structural and vibrational data can be dependent on the particular solvent reaction field model used and other calculational details (see also ref 53), our results point to the necessity to consider (at least) the first hydration shell around acetate explicitly. This can explain a mismatch between the calculated C−O stretching frequencies and those observed in the condensed phase of sodium carboxylates in a recent study by Oomens and Steill,[9] whereas their calculated values agreed with the experimental frequencies for the isolated carboxylates in the gas phase. In their work, the calculated frequencies have been scaled to account for anharmonic effects and other systematic deficiencies originated from the chosen level of calculation

Table 6. Selected Structural Parameters and Frequencies (in cm^{-1}) for the Optimized Acetate−Water Clusters (B3LYP/aug-cc-pVDZ)

n^a	r(C−O), Å	∠O−C−O, deg	r(C−C), Å	ν_{as}(C−O)	ν_s(C−O)^b	ν(C−C)	δ(OCO)^b
0^c	1.262	128.6	1.560	1634	1348	865	624
0 (CPCM)^d	1.269	124.7	1.531	1503	1392	897	639
1	1.265	127.5	1.547	1615	1373	888	638
2	1.258, 1.270	127.6	1.542	1621	1390	888	640
3	1.259, 1.273	126.6	1.536	1608	1397	921	640
4	1.260, 1.272	125.0	1.530	1612	1418	943	659
5	1.262, 1.274	124.5	1.526	1599	1420	931	660
6	1.259, 1.278	124.5	1.524	1598	1420	925	665
1 (Na)^e	1.268, 1.273	125.1	1.522	1573	1422	923	679
2 (Na)^e	1.260, 1.277	124.2	1.521	1581	1430	934	678
3 (Na)^e	1.262, 1.273	126.3	1.523	1568	1427	923	659
4 (Na)^e	1.260, 1.278	124.0	1.520	1589	1434	920	664

^{a} Number of water molecules in the cluster (see Figure 3). ^{b} With the significant contribution from the C−C stretching and other modes. ^{c} Isolated acetate anion in a vacuum. ^{d} Isolated acetate anion in a dielectric medium (water). ^{e} Complex with the sodium cation and water molecules (see Figure 3).
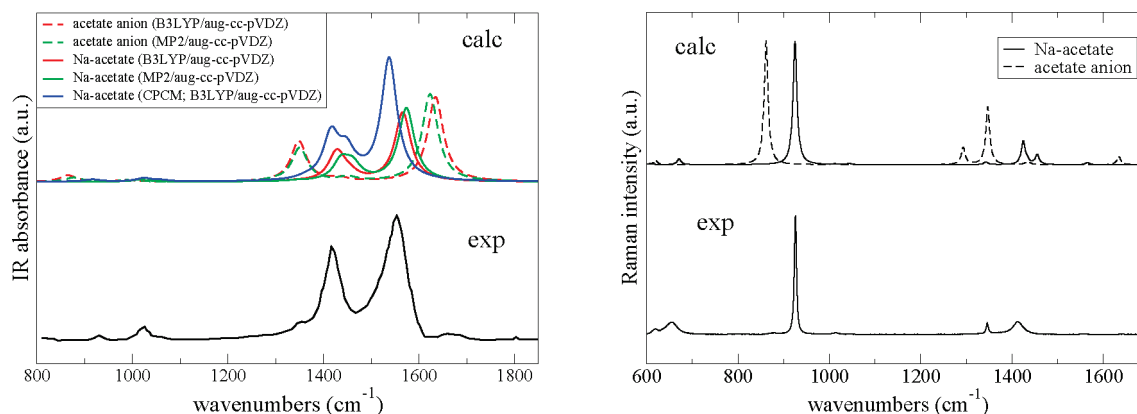

Figure 4. Left: Calculated harmonic IR spectra for the free acetate anion and its sodium bidentate complex. Experimental spectrum for the aqueous solution of sodium acetate is taken from ref 46. Right: Simulated Raman spectra for the acetate anion and its sodium bidentate complex (B3LYP/aug-cc-pVDZ). The wavelength of the diode laser (785 nm), as used in the experiment, is assumed, and line shapes are modeled with the Lorentz function with a resolution of 10 cm^{-1} at room temperature. The experimental Raman spectrum of solid sodium acetate is taken from ref 67.

(see also ref 8). Oomens and Steill,[9] however, have pointed to the counterion binding as a possible reason for such a discrepancy. As can be seen from our results, upon bidentate metal coordination, the symmetric carboxylate mode is affected much stronger than the asymmetric one. As far as the formate anion is concerned, the large discrepancy between the calculated (free)[32] and experimental (aqueous solution) vibrational frequencies can be explained by the influence of the counterion (sodium) present in the condensed phase. The most striking mismatch was observed for the C−H stretching band, where the calculated values of 2657 cm^{-1} (CCSD(T)/aug-cc-pVQZ),[34] 2597 cm^{-1} (MP4/6-311++G(d, p)),[32] and 2573 cm^{-1} (B3LYP/aug-cc-pVTZ, this work) for the free formate significantly underestimated the values of 2821 cm^{-1} observed in an aqueous solution of its sodium salt[52] and 2830 cm^{-1} measured for polycrystalline sodium formate.[45] After correcting the calculated value for anharmonicity (ca. 200 cm^{-1}), Krekeler et al.[34] obtained a very good agreement between their frequency (2441 cm^{-1}) and the experimental gas-phase value (2456 cm^{-1}).[7] Our calculated value (2926 cm^{-1}) for sodium formate (see Figure 2), being corrected for anharmonicity (2748 cm^{-1}), is much closer to the experimental value in aqueous solution mentioned above. This blue-shift of the C−H stretching band is

solely due to the significant shortening of the C−H bond distance upon coordination to a metal cation (1.105 Å) as compared with the value in the free formate anion (1.135 Å). This value is in perfect agreement with the low-temperature (120 K) neutron diffraction study on sodium formate of Fuess et al. (1.1004(7) Å).[39] Contrary to the statement of Dixon et al.,[33] poor agreement is observed between the experimental solid-state frequencies[45] and their calculated values (CCSD(T)/aug-cc-pVDZ) for the C−H stretching, C−O symmetric stretching, and OCO bending modes since the influence of the counterion (sodium) was ignored.

In Figure 4, we compare the calculated IR and Raman spectra for the free acetate anion and its sodium bidentate complex at different levels of theory. A reduced splitting between the carboxylate asymmetric and symmetric stretching bands is clearly seen for the sodium acetate in the calculated IR spectrum, which reproduces nicely the experimental IR pattern.[46,50]

It has been shown by Raman spectroscopy[49,50,68] that the C−C stretching mode in acetate, being the most intense band, is a very sensitive probe of the local chemical environment around the C−C bond. Our calculations (B3LYP/aug-cc-pVDZ) give for this mode a value of 928 cm^{-1} (sodium

Table 7. Projected Force Constants for the Carboxylate Moiety

parameter[a]	free carboxylate			sodium carboxylate		
	formate	acetate	benzoate	formate	acetate	benzoate
C−O str, mdyn/Å	9.83	9.62	9.77	9.51	9.11	9.04
C−R str, mdyn/Å	3.74	3.27	3.48	4.75	4.14	4.44
O−C−O bend, mdyn Å/rad²	0.80	0.86	0.90	1.01	1.08	1.11
O−C−R bend, mdyn Å/rad²	0.54	0.72	0.83	0.57	0.74	0.85
O−C−O wag, mdyn Å/rad²	0.06	0.07	0.07	0.06	0.07	0.07
C−O/C−O str−str, mdyn/Å	1.47	1.52	1.47	1.58	1.59	1.52
C−O/C−R str−str, mdyn/Å	0.63	0.71	0.78	0.40	0.55	0.65
C−O/O−C−O str−bend, mdyn/rad	0.19	0.20	0.21	0.31	0.32	0.34
C−O/O−C−R str−bend, mdyn/rad	0.16	0.20	0.20	0.09	0.13	0.12
C−R/O−C−R str−bend, mdyn/rad	0.12	0.18	0.24	0.08	0.15	0.20
(O, O)C−R tors, mdyn Å/rad²		0.007	0.06		0.008	0.08

[a] Abbreviations used: "str" for stretch; "bend" for in-plane angle bend; "wag" for out-of-plane angle bend; "str−str" and "str−bend" for stretch−stretch and stretch−bend interactions, respectively; and "tors" for COO twisting.

acetate), which deviates appreciably from that found for the free acetate anion (865 cm^{-1}). This is in accord with a much shorter C−C bond distance in the bidentate complex with a sodium cation (1.521 Å) as compared to the value in the isolated acetate (1.560 Å). The simulated Raman spectra for the acetate anion and its sodium bidentate complex, shown in Figure 4, are clearly distinguishable, with the latter being in qualitative agreement with the available experimental data.[48,67,69]

The interpretation of the Raman bands observed recently from a single supersaturated droplet of sodium acetate by Wang et al.[70] should probably be reinterpreted. Their band assignment (see Table 1) is in perfect agreement with our results for the sodium bidentate complex if one assumes that, both in dilute solution and in a supersaturated droplet, the acetate group is appreciably perturbed by a counterion.

Having demonstrated the reliability of the chosen computational scheme, we are now in a position to develop the force field parameters, which can be used in molecular simulations where the main focus is on obtaining both accurate geometries and dynamics (vibrational frequencies) of the carboxylate moiety. The derivation of the FF parameters for the formate anion is particularly simple, since, in accord with the MM3 methodology, no van der Waals or electrostatic terms need to be considered (no 1−4 interactions). This is in contrast with the statement of Kirschner et al.,[4] where difficulties have been encountered when developing parameters for this particular anion. As already mentioned above, this is solely due to the absence of some important cross-terms (stretch−stretch) in the original MM3 functional form.

The calculated Hessian matrix in Cartesian coordinates (B3LYP/aug-cc-pVTZ) was projected into the redundant set of internal coordinates at the trigonal-planar C$_{carb}$ center (three bond stretchings, three in-plane and three out-of-plane angle bendings) both for the free carboxylate ion and for its sodium bidentate complex. The number of parameters to be refined was nine (diagonal terms) and 14 (diagonal and off-diagonal terms).

Besides the influence of the substituent group (R = H, Me, or Ph) clearly seen from the data presented in Table 7, the perturbation of the force constants caused by sodium chelation is also evident. On going from the free carboxylate ion to its complex with the sodium cation, the C−O bond

stretch force constants decrease by ~3% (formate), ~5% (acetate), and ~7% (benzoate), whereas the O−C−O angle in-plane bending force constants increase by ~20%. On the basis of a detailed analysis of the normal modes, the influence of the sodium cation can be described as follows. Since the mode, commonly referred to as "asymmetric" (or sometimes also as "antisymmetric") stretching, is solely due to the out-of-phase stretching of the two C−O bonds, its frequency should decrease on complexation with the cation simply because the C−O stretch force constant decreases (see Table 7). At the same time, the second mode, commonly referred to as "symmetric" stretching, is actually a mixture of the in-phase stretching of the two C−O bonds and the C−C stretching (in acetate and benzoate) with a small contribution from the O−C−O angle bending. Since both the C−C stretch and the O−C−O angle bend force constants increase on complexation (see Table 7) and, in addition, the contribution from the C−C stretch to this mode increases (from 16% in acetate and benzoate to 26% in sodium acetate and benzoate, on the basis of the potential energy distribution), their overall effect overwhelms the decrease of the C−O stretch force constant, resulting in a higher frequency as compared to that of the free carboxylate anion (compare the data in Tables 3 and 4).

We also found that the rotation around the C$_{carb}$−C bond in the acetate is essentially free, which is in agreement with previous theoretical[4,53] and experimental (spectroscopic)[49] studies. The torsional force constant given by Kakihana et al.[47] is probably too large, which is due to the overestimation of the barrier to methyl group rotation (2.3 kcal/mol)[71] and that of the corresponding torsional frequency derived therefrom used in the fit. A recent neutron scattering study of Moreno et al.[72] gave a value of 0.95 kcal/mol for the rotational barrier in anhydrous sodium acetate. This latter value is comparable with that found earlier by Montjoie and Müller-Warmuth[73] (0.77 kcal/mol) from their analysis of the correlation between NMR spin−lattice relaxation and neutron-scattering data on anhydrous sodium acetate. The rotational barrier in the benzoate, on the other hand, is appreciable (vide supra).

Introducing two new atomic types, C$_{carb}$ and O$_{carb}$, and taking minor modifications into account (see eqs 1 and 2), the force field parameters presented in Table 8 are suitable for use within the MM3(2000) force field. As in our previous

2830 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Tafipolsky and Schmid

Table 8. Additional MM3(2000) Force Field Parameters for the Carboxylate Moiety

Bond Stretches		
atom types	reference distance, Å	force parameter, mdyn/Å
$C_{carb}-O_{carb}$	1.252 (1.264)[a]	9.8 (9.3)
$C_{carb}-H$	1.13 (1.11)	3.7 (4.6)
$C_{carb}-C(sp^3)$	1.542 (1.514)	3.6 (4.2)
$C_{carb}-C_{ph}$	1.531 (1.481)	3.8 (5.1)

In-Plane Angle Bending		
	reference angle, deg	force parameter, mdynÅ/rad^2
$O_{carb}-C_{carb}-O_{carb}$	130.0 (124.0)	1.4 (1.9)
$O_{carb}-C_{carb}-H$	114.8 (114.8)	0.7 (0.7)
$O_{carb}-C_{carb}-C(sp^3)$	117.3 (117.8)	0.9 (0.9)
$O_{carb}-C_{carb}-C_{ph}$	115.4 (121.0)	1.0 (0.8)
Out-of-Plane Angle Bending		
$(R-)C_{carb}-O_{carb}$[b]	0.0	1.7 (1.7)
Stretch−Stretch (mdyn/Å)[c]		
$C_{carb}-O_{carb}/C_{carb}-C_{carb}$		1.5 (1.6)
$C_{carb}-O_{carb}/C_{carb}-R$		0.7 (0.5)
Stretch−Bend (mdyn/rad)		
$C_{carb}-O_{carb}/O_{carb}-C_{carb}-O_{carb}$		0.6 (0.8)
$C_{carb}-O_{carb}/O_{carb}-C_{carb}-R$		0.6 (0.5)
$C_{carb}-R/O_{carb}-C_{carb}-R$		0.5 (0.4)

[a] Values for a carboxylate moiety perturbed by bidentate coordination of sodium are given in parentheses. [b] All three Wilson angles at a trigonal center are assigned the same force constant. [c] Due to the error in the implementation of eq 1, the values cited in Table 9 of ref 1 should be exactly 2 times smaller.

work,[1] we decided to use effective atomic point charges to describe electrostatic interactions instead of bond dipoles normally used within the MM3 formalism. This is done in order to facilitate the applicability of derived FF parameters within other MM implementations. It turned out, however, that the electrostatic interactions can be ignored since, as we found in our preliminary tests, they had no influence on the parameters we are interested in except for the $H-C-C_{carb}-O_{carb}$ and $C_{ph}-C_{ph}-C_{carb}-O_{carb}$ torsional parameters for acetate and benzoate, respectively. Since the values for the torsional parameters are dependent (especially for benzoate) on the particular electrostatic model used (point charges, bond dipoles, etc.), they are not included in Table 8. The standard MM3(2000) van der Waals parameters were employed without further refinement, since we focus here on intramolecular interactions, and the force field parameters were adjusted with a genetic algorithm used in our previous work.[1]

When the atomic point charges fitted to reproduce the electrostatic potential (+0.8 and −0.8 for C_{carb} and O_{carb}, respectively, 0.0 for the $C_{ph}(-C_{carb})$, and average values of −0.12 and +0.12 for $C_{ph}(-H)$ and H, respectively) were used, and assuming a single 2-fold Fourier term $(1/2V_2(1 - \cos2\phi))$ for the $C_{ph}-C_{ph}-C_{carb}-O_{carb}$ twist angle (ϕ), the parameters V_2 for the isolated benzoate anion and its complex with the sodium cation were refined to be 1.4 and 1.7 kcal/mol, respectively. With these parameters at hand, we reproduce the calculated (B3LYP/aug-cc-pVTZ) lowest (torsional) mode of the isolated benzoate anion (55 cm^{-1}) and its complex with the sodium cation (67 cm^{-1}) within

4 cm^{-1}. If the electrostatic interactions are ignored completely, we arrived at the values of 2.0 and 2.3 kcal/mol, respectively, for the V_2 parameter.

Using the derived FF parameters, the calculated vibrational frequencies for all three carboxylates, characteristic for the carboxylate moiety (symmetric and asymmetric C−O stretchings and O−C−O in-plane bending), are very close to the reference values (B3LYP/aug-cc-pVTZ) with a root-mean-squared deviation of less than 14 cm^{-1}. This is particularly astonishing since, in addition to the diagonal FF parameters, only a few cross-terms (stretch−stretch and stretch−bend) listed in Table 8 were explicitly optimized. Interestingly, due to the redundancy relationship among the three valence angles around the trigonal-planar C_{carb} atom, the values for other cross-terms (see Figure 1) are reproduced automatically. We should note that the importance of the CO/CO stretch−stretch cross-term to reproduce the stretching frequencies of the carboxylate moiety was established by spectroscopists quite a long time ago. Probably, one of the first estimates was due to Jones and McLaren[74] from the analysis of IR spectra of solid anhydrous sodium acetate (1.7 mdyn/Å), where the structural parameters for the carboxylate group (C−O bond distances and O−C−O bond angle) were taken from the crystal structure of sodium formate.[75] The values given by Spinner[76] (1.6 mdyn/Å) and Kidd and Mantsch[45] (1.6 mdyn/Å) for formate, by Beckmann et al.[77] (1.3 mdyn/Å) and Kakihana et al.[47] for acetate (2.0 mdyn/Å), and by Ernstbrunner et al.[78] for 4-nitrobenzoate (1.5 mdyn/Å) are significant and are very close to the value derived in the present work. We should also mention the large value for this particular stretch−stretch cross-term obtained by Lii[6] for formic acid (1.3 mdyn/Å). Further, our value for the CO/OCO stretch−bend interaction (in mdyn/rad) is also very similar to those available in the literature (0.3[45] for formate, 0.4[77] for acetate, and 0.6[78] for 4-nitrobenzoate). Relevant to our compilation (see Table 8) are the parameters used to describe the carboxylate group in another widely used MMFF94 force field.[79] There, the values of 9.756 mdyn/Å, 1.181 mdyn Å/rad^2, and 0.652 mdyn/rad for the CO bond stretch, OCO angle bend, and CO/OCO stretch−bend interaction, respectively, are used. These values are strikingly similar to those derived in our study. However, no stretch−stretch cross-term is used for this moiety in the MMFF94 force field.

We should mention that, among numerous previous attempts to reproduce the vibrational spectra of carboxylates by refining the corresponding force constants, a particular success of Kakihana et al.[47] in their study of the acetate anion can be attributed mainly to the use of many important cross-terms in the valence force field (actually, due to the redundancy relationship around the trigonal-planar C_{carb} center, even less cross-terms are needed to parametrize explicitly). Their effective force constants, refined against 82 observed frequencies of six isotopomers of sodium acetate, have implicitly incorporated the influence of the metal cation on vibrational spectra. In another earlier work, Dasgupta and Goddard[3] combined the Hessian from ab initio calculations with the structural and spectroscopic data from experimental data to generate the force field parameters for
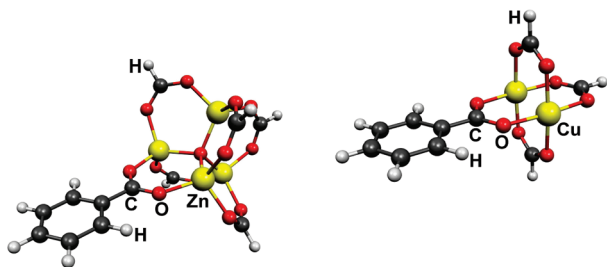
A Consistent Force Field for the Carboxylate Group

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2831**

Figure 5. Two motifs encountered in MOFs based on a basic metal carboxylate, $Zn_4O(O_2CH)_5$-benzoate (left), and a binuclear metal ("paddle-wheel") carboxylate, $Cu_2(O_2CH)_3$-benzoate (right), building units.

the formate anion. On the basis of their recommended parameters (see Table 2, column labeled as "HTX" in ref 3), we were able to reproduce their large deviation ($\sim$90 cm^{-1}) for the lowest-frequency OCO bend (see Table 10, column labeled as "HTX" in ref 3). The reason for this discrepancy could be their smaller values for the OCO bending and CO/OCO stretch−bend interaction force constants as compared to our results.

Our motivation in this work was primarily to develop reliable FF parameters for the carboxylate group directly from the quantum chemical calculations, which can be used in molecular simulations of MOF materials, where this particular functionality is used to link various metal−oxide clusters together, resulting in a 3D porous framework.[1] Among the variety of linkers used so far, one ubiquitous type consists of the carboxylate functionality conjugated with the phenyl moiety,[80] for which benzoate can serve as a model system. In modeling the frameworks we are interested in, one important parameter is the barrier to internal rotation of the carboxylate group with respect to the phenyl ring. In contrast to the bidentate coordination of an alkali metal (see Figure 2), metal−oxide clusters usually bind in a bridging fashion exemplified in Figure 5 by two representative motifs encountered in MOFs based on basic metal carboxylate and dinuclear metal tetracarboxylate ("paddle-wheel" motif) building units.

In our calculations, the use of the benzoate moiety instead of a more realistic linker encountered in various MOFs, such as benzene-1,4-dicarboxylate (terephthalate) in MOF-5[81] or benzene-1,3,5-tricarboxylate (trimesate) in CuBTC,[82] needs to be justified. Moreover, replacing the bridging metal−oxide cluster (as shown in Figure 5) with the bidentate coordination of an alkali metal can in principle modify the barrier to rotation as well. To this end, the geometry of an "orthogonal" conformer was optimized, constrained to have the carboxylate group rotated by 90° with respect to the phenyl ring. The torsional barrier was calculated as the energy difference between the global minimum (all atoms are in one plane) and the orthogonal conformer at the B3LYP level of theory using the correlation consistent basis sets of improved quality (from double- to triple-ζ).

The results presented in Table 9 allow some important conclusions to be drawn. First, the augmentation of basis sets with diffuse functions (marked with the prefix "aug") has a large effect on the barrier, especially for the free benzoate anion. Second, going to a larger basis set lowers

Table 9. Calculated Barriers to Rotation of the Carboxylate Group (kcal/mol) for Model Systems

	cc-pVDZ	aug-cc-pVDZ	cc-pVTZ	aug-cc-pVTZ[a]
benzoate anion	8.2 (8.6)[b]	4.0 (4.5)	5.2 (6.0)	3.8 (4.6)
Li-benzoate	8.1 (7.7)	6.8 (6.3)	7.1 (6.9)	6.8
Na-benzoate	7.9 (7.5)	6.2 (5.9)	6.4 (6.4)	6.0
$Zn_4O(O_2CH)_5$-benzoate	8.4	6.9	7.0	
$Cu_2(O_2CH)_3$-benzoate[c]	7.5	6.3	6.4	

[a] A basis set without diffuse functions (cc-pVTZ) was used for all H atoms. [b] Values calculated at the MP2(frozen core) level of theory are given in parentheses. [c] Triplet electronic state.

Table 10. Calculated Barriers to Rotation of the Carboxylate Group (kcal/mol) for Terephthalate and Trimesate Anions and Their Salts[a]

	1	2	3
terephthalate anion	4.2	9.2	
Li$_2$-terephthalate	6.4	13.5	
Na$_2$-terephthalate	5.8	12.5	
trimesate anion	2.0	4.9	8.4
Li$_3$-trimesate	6.1	12.8	19.9
Na$_3$-trimesate	5.4	11.4	17.9

[a] Numbers (1, 2, or 3) indicate the number of COO groups orthogonal to the benzene plane.

the barrier. Third, a triple-ζ basis set (aug-cc-pVTZ) gives similar results as a much more economical double-ζ basis set (aug-cc-pVDZ). Therefore, the latter basis set can be recommended. We also note that the barriers calculated at the MP2 level of theory using the aug-cc-pVDZ basis set are close to the values obtained at the DFT/B3LYP level. Interestingly, the rotational barriers calculated for bidentate lithium and sodium benzoates are similar to those of bridging zinc and copper benzoates (see Figure 5), respectively. It is also evident that coordination with metals (or metal−oxides) makes the rotational barrier higher in comparison with that in the free carboxylate. To the best of our knowledge, there is no systematic study on the barrier to rotation of the carboxylate group conjugated with aromatic rings and its dependence on the mode of coordination to metal cations. Rakitin and Pack[83] have calculated a barrier of 3.4 kcal/mol for the free benzoate anion (MP2/STO-3G), whereas Woo et al.[84] gave a value of $\sim$4.5 kcal/mol for the p-methylbenzoate anion (B3LYP/6-31+G(d, p)).

To get some more insight into the rotational flexibility of the carboxylate group conjugated with the benzene ring, we have calculated (B3LYP/aug-cc-pVDZ) the most important stationary points on potential energy hypersurfaces of terephthalate and trimesate, both as free anions and as chelated (bidentate) complexes with alkali metals (lithium and sodium). The results for rotational isomers (conformers) are presented in Table 10, where the plane of one or two (terephthalate) and one, two, or three (trimesate) COO groups can be orthogonal to the plane of the benzene ring. The energies are given with respect to the global minimum structure, which corresponds to a planar arrangement of all atoms.

The most evident trend seen from these data is that the energy penalty for two (terephthalate) and for two and three (trimesate) orthogonal COO groups is not quite additive; that

2832 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Tafipolsky and Schmid

is, the energy cost increases with more carboxylate groups rotated out of the benzene plane. Comparing the calculated barriers (B3LYP/aug-cc-pVDZ) in dilithium (13.5 kcal/mol) and disodium (12.5 kcal/mol) terephthalate (in this case, an "orthogonal" conformer corresponds to the geometry where both COO groups are rotated by 90° with respect to the phenylene ring) with the doubled value calculated for lithium ($2 \times 6.8 = 13.6$ kcal/mol) and sodium ($2 \times 6.2 = 12.4$ kcal/mol) benzoates (see Table 9), one can conclude that the energy penalty arising from the rotation of the carboxylate group out of the benzene plane is additive to a large degree. This observation justifies the use of benzoate as a model system for more complicated benzene polycarboxylates such as terephthalate or trimesate. We also note that the calculated torsional barrier is much closer to a recently determined experimental value of 11 ± 2.0 kcal/mol[85] as compared to our previous estimation (~16 kcal/mol; B3LYP/cc-pVDZ),[1] which is due to a larger basis set used in this work (augmented with diffuse basis functions). For some recent estimations of the rotational barriers, see also ref 86.

4. Conclusions

In summary, we have clarified the role played by a counterion in the interpretation of the vibrational spectra of the carboxylate group in the condensed phase (aqueous solution and solid state) using formate, acetate, and benzoate and their sodium bidentate complexes as examples. We have shown that, with the help of such simplified models, it is possible to explain the most salient features observed in spectral regions characteristic of the carboxylate vibrations. In light of new experimental data corroborated by the calculations, the notion of a "free" (isolated) carboxylate needs to be revised. Even in dilute (aqueous) solutions, a carboxylate group can be perturbed by a counterion quite appreciably. We argue that the direct or indirect influence of a counterion was largely underestimated in the past and cannot be ignored. On the basis of the results of ab initio calculations, the changes in the force field of the carboxylate moiety upon metal ion coordination are analyzed in detail. With the use of density functional theory along with a large basis set augmented with diffuse basis functions, the molecular-mechanical force field parameters are developed, aimed at molecular simulations of large biomolecules, such as amino acids, as well as that of a broad class of metal−organic coordination polymers where polydentate carboxylates are used as organic linkers.[1] The use of a complete (redundant) set of internal coordinates is not only a straightforward and natural approach to developing the FF parameters but, in addition, can reduce the number of cross-terms necessary to reliably predict the vibrational frequencies within the MM approach, thus avoiding complications arising from utilizing some additional terms, like the Urey−Bradley term.[54] In addition to the CO/OCO stretch−bend cross-term, the importance of the CO/CO stretch−stretch interaction in the carboxylate moiety is emphasized. A similar approach can be applied for molecular-mechanical force field derivation for other important anions like phosphate, nitrate, sulfate, carbonate, and so forth. We can anticipate that the presented analysis and derived parameters will be useful in the modeling of solid metal carboxylates per se,[87] where not only the structure but, in addition, accurate dynamics are of interest. Simulations of transition metal carboxylates within the MM approach[88] can also benefit from our findings. With the use of the results of this study, the influence of a particular mode of metal coordination to a carboxylate group, as manifests itself in the fingerprint region of vibrational spectra, can be modeled reliably and accurately, helping to interpret IR spectra of various MOFs. Last but not least, molecular simulations of amino acids in solution (e.g., as zwitterions in water) can be performed by taking full internal flexibility of the carboxylate moiety in the force field into account.

References

(1) Tafipolsky, M.; Schmid, R. *J. Phys. Chem. B* **2009**, *113*, 1341.

(2) Maple, J. R.; Dinur, U.; Hagler, A. T. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350.

(3) Dasgupta, S.; Goddard, W. A. *J. Chem. Phys.* **1989**, *90*, 7207.

(4) Kirschner, K. N.; Lewin, A. H.; Bowen, J. P. *J. Comput. Chem.* **2003**, *24*, 111.

(5) Nevins, N.; Chen, K.-S.; Allinger, N. L. *J. Comput. Chem.* **1996**, *17*, 669.

(6) Lii, J.-H. *J. Phys. Chem. A* **2002**, *106*, 8667.

(7) Forney, D.; Jacox, M. E.; Thompson, W. E. *J. Chem. Phys.* **2003**, *119*, 10814.

(8) Oomens, J.; Steill, J. D. *J. Phys. Chem. A* **2008**, *112*, 3281.

(9) Steill, J. D.; Oomens, J. *J. Phys. Chem. A* **2009**, *113*, 4941.

(10) Deacon, G. B.; Phillips, R. J. *Coord. Chem. Rev.* **1980**, *33*, 227.

(11) Lewandowski, W.; Kalinowska, M.; Lewandowska, H. *J. Inorg. Biochem.* **2005**, *99*, 1407.

(12) Nara, M.; Torii, H.; Tasumi, M. *J. Phys. Chem.* **1996**, *100*, 19812.

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

(14) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

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

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

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

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

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

(20) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **1999**, *110*, 7667.

(21) Figgen, D.; Rauhut, G.; Dolg, M.; Stoll, H. *Chem. Phys.* **2005**, *311*, 227.

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

(23) Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571.

(24) Schuchardt, K. L.; Didler, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem.*

Inf. Model. **2007**, *47*, 1045. Accessible at https://bse.pnl.gov/bse/portal (accessed August 2009).

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.04; Gaussian, Inc.: Pittsburgh, PA, 2003.

(26) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.

(27) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(28) Ponder, J. W.; Ren, P.; Pappu, R. V.; Hart, R. K.; Hodgson, M. E.; Cistola, D. P.; Kundrot, C. E.; Richards, F. M. *TINKER*, version 4.2; Washington University School of Medicine: St. Louis, MO, 2004 Avvailable at http://dasher.wustl.edu/tinker/ (accessed August 2009).

(29) Maple, J. R.; Hwang, M. J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comput. Chem.* **1994**, *15*, 162.

(30) Halgren, T. A. *THEOCHEM* **1988**, *163*, 431.

(31) Palmo, K.; Pietila, L.-O.; Krimm, S. *J. Comput. Chem.* **1992**, *13*, 1142.

(32) Magalhaes, A. L.; Madail, S. R. R. S.; Ramos, M. J. *Theor. Chem. Acc.* **2000**, *105*, 68.

(33) Dixon, D. A.; Feller, D.; Francisco, J. S. *J. Phys. Chem. A* **2003**, *107*, 186.

(34) Krekeler, C.; Mladenovic, M.; Botschwina, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 882.

(35) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.

(36) Carrell, C. J.; Carrell, H. L.; Erlebacher, J.; Glusker, J. P. *J. Am. Chem. Soc.* **1988**, *110*, 8651.

(37) Hsu, L. Y.; Nordman, C. E. *Acta Crystallogr., Sect. C* **1983**, *39*, 690.

(38) Markila, P. L.; Rettig, S. J.; Trotter, J. *Acta Crystallogr., Sect. B* **1975**, *31*, 2927.

(39) Fuess, H.; Bats, J. W.; Dannohl, H.; Meyer, H.; Schweig, A. *Acta Crystallogr., Sect. B* **1982**, *38*, 736.

(40) Glusker, J. P. *Acta Crystallogr., Sect. D* **1995**, *51*, 418.

(41) Aziz, E. F.; Ottosson, N.; Eisebitt, S.; Eberhardt, W.; Jagoda-Cwiklik, B.; Vacha, R.; Jungwirth, P.; Winter, B. *J. Phys. Chem. B* **2008**, *112*, 12567.

(42) Wei, K. T.; Ward, D. L. *Acta Crystallogr., Sect. B* **1977**, *33*, 522.

(43) Efremov, V. A.; Endeladze, N. O.; Agre, V. M.; Trunov, V. K. *J. Struct. Chem.* **1986**, *27*, 498.

(44) Helmholdt, R. B.; Sonneveld, E. J.; Schenk, H. *Z. Kristallogr.* **1998**, *213*, 596.

(45) Kidd, K. G.; Mantsch, H. H. *J. Mol. Spectrosc.* **1981**, *85*, 375.

(46) Pike, P. R.; Sworan, P. A.; Cabaniss, S. E. *Anal. Chim. Acta* **1993**, *280*, 253.

(47) Kakihana, M.; Kotaka, M.; Okamoto, M. *J. Phys. Chem.* **1983**, *87*, 2526.

(48) Frost, R. L.; Kloprogge, J. T. *J. Mol. Struct.* **2000**, *526*, 131.

(49) Bickley, R. I.; Edwards, H. G. M.; Rose, S. J.; Gustar, R. *J. Mol. Struct.* **1990**, *238*, 15.

(50) Quiles, F.; Burneau, A. *Vib. Spectrosc.* **1998**, *16*, 105.

(51) Green, J. H. S. *Spectrochim. Acta, Part A* **1977**, *33*, 575.

(52) Spinner, E. *J. Chem. Soc. B* **1967**, 879.

(53) Keresztury, G.; Istvan, K.; Sundius, T. *J. Phys. Chem. A* **2005**, *109*, 7938.

(54) Meziane-Tani, M.; Lagant, P.; Semmoud, A.; Vergoten, G. *J. Phys. Chem. A* **2006**, *110*, 11359.

(55) Remko, M.; Van Duijnen, P. T.; von der Lieth, C. W. *THEOCHEM* **2007**, *814*, 119.

(56) Markham, G. D.; Bock, C. L.; Bock, C. W. *Struct. Chem.* **1997**, *8*, 293.

(57) Payaka, A.; Tongraar, A.; Rode, B. M. *J. Phys. Chem. A* **2009**, *113*, 3291.

(58) Leung, K.; Rempe, S. B. *J. Am. Chem. Soc.* **2004**, *126*, 344.

(59) Kameda, Y.; Sasaki, M.; Yaegashi, M.; Tsuji, K.; Oomori, S.; Hino, S.; Usuki, T. *J. Solution Chem.* **2004**, *33*, 733.

(60) da Silva, E. F.; Svendsen, H. F.; Merz, K. M. *J. Phys. Chem. A* **2009**, *113*, 6404.

(61) Gojlo, E.; Smiechowski, M.; Panuszko, A.; Stangret, J. *J. Phys. Chem. B* **2009**, *113*, 8128.

(62) Myshakin, E. M.; Jordan, K. D.; Sibert, E. L.; Johnson, M. A. *J. Chem. Phys.* **2003**, *119*, 10138.

(63) Michaux, C.; Wouters, J.; Perpete, E. A.; Jacquemin, D. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 632.

(64) Epstein, L. M.; Saitkulova, L. N.; Shubina, E. S. *J. Mol. Struct.* **1992**, *270*, 325.

(65) Wang, X.-B.; Jagoda-Cwiklik, B.; Chi, C.; Xing, X.-P.; Zhou, M.; Jungwirth, P.; Wang, L.-S. *Chem. Phys. Lett.* **2009**, *477*, 41.

(66) Park, J. Y.; Woon, D. E. *Astrophys. J.* **2006**, *648*, 1285.

(67) de Veij, M.; Vandenabeele, P.; De Beer, T.; Remonc, J. P.; Moens, L. *J. Raman Spectrosc.* **2009**, *40*, 297.

(68) Nickolov, Z.; Ivanov, I.; Georgiev, G.; Stoilova, D. *J. Mol. Struct.* **1996**, *377*, 13.

(69) Ito, K.; Bernstein, H. J. *Can. J. Chem.* **1956**, *34*, 170.

(70) Wang, L. Y.; Zhang, Y. H.; Zhao, L. J. *J. Phys. Chem. A* **2005**, *109*, 609.

(71) Kakihana, M.; Kotaka, M.; Okamoto, M. *J. Phys. Chem.* **1983**, *87*, 3510.

(72) Moreno, A. J.; Alegria, A.; Colmenero, J.; Frick, B. *Appl. Phys. A: Mater. Sci. Process.* **2002**, *74*, S135.

(73) Montjoie, A.-S.; Müller-Warmuth, W. *Z. Naturforsch. A* **1985**, *40*, 596.

(74) Jones, L. H.; McLaren, E. *J. Chem. Phys.* **1954**, *22*, 1796.

(75) Zachariasen, W. H. *J. Am. Chem. Soc.* **1940**, *62*, 1011.

(76) Spinner, E. *J. Chem. Soc. B* **1967**, 874.

(77) Beckmann, L.; Gutjahr, L.; Mecke, R. *Spectrochim. Acta* **1964**, *20*, 1295.

(78) Ernstbrunner, E. E.; Girling, R. B.; Hester, R. E. *J. Chem. Soc., Faraday Trans.* **1978**, *74*, 1540.

(79) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490. The MMFF94 parameter files can be accessed at http://www.mrw.interscience.wiley.com/suppmat/0192-8651/suppmat/aid.0176.html (accessed 06/10/2009).

(80) Rowsell, J. L. C.; Yaghi, O. M. *Microporous Mesoporous Mater.* **2004**, *73*, 3.

(81) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. *Nature* **1999**, *402*, 276.

(82) Chui, S. S. Y.; Lo, S. M. F.; Charmant, J. P. H.; Orpen, A. G.; Williams, I. D. *Science* **1999**, *283*, 1148.

(83) Rakitin, A. R.; Pack, G. R. *Langmuir* **2005**, *21*, 837.

(84) Woo, H.-K.; Wang, X.-B.; Kiran, B.; Wang, L.-S. *J. Phys. Chem. A* **2005**, *109*, 11395.

(85) Gould, S. L.; Tranchemontagne, D.; Yaghi, O. M.; Garcia-Garibay, M. A. *J. Am. Chem. Soc.* **2008**, *130*, 3246.

(86) Winston, E. B.; Lowell, P. J.; Vacek, J.; Chocholousova, J.; Michl, J.; Price, J. C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5188.

(87) Barreto, L. S.; Mort, K. A.; Jackson, R. A.; Alves, O. L. *J. Phys.: Condens. Matter* **2000**, *12*, 9389.

(88) Deeth, R. J. *Inorg. Chem.* **2008**, *47*, 6711.

JCTC Journal of Chemical Theory and Computation

Performance of DFT Methods in the Calculation of Optical Spectra of TCF-Chromophores

Jan Andzelm,*,[†] Berend C. Rinderspacher,[†] Adam Rawlett,[†] Joseph Dougherty,[†] Roi Baer,[‡] and Niranjan Govind[§]

U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland 21005, Institute of Chemistry and Fritz Haber Center for Molecular Dynamics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel, and William R. Wiley Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352

Abstract: We present electronic structure calculations of the ultraviolet/visible (UV−vis) spectra of highly active push−pull chromophores containing the tricyanofuran (TCF) acceptor group. In particular, we have applied the recently developed long-range corrected Baer-Neuhauser-Livshits (BNL) exchange-correlation functional. The performance of this functional compares favorably with other density functional theory (DFT) approaches, including the CAM-B3LYP functional. The accuracy of UV−vis results for these molecules is best at low values of attenuation parameters (γ) for both BNL and CAM-B3LYP functionals. The optimal value of γ is different for the charge-transfer (CT) and $\pi-\pi^*$ excitations. The BNL and PBE0 exchange correlation functionals capture the CT states particularly well, while the $\pi-\pi^*$ excitations are less accurate and system dependent. Chromophore conformations, which considerably affect the molecular hyperpolarizability, do not significantly influence the UV−vis spectra on average. As expected, the color of chromophores is a sensitive function of modifications to its conjugated framework and is not significantly affected by increasing aliphatic chain length linking a chromophore to a polymer. For selected push−pull aryl-chromophores, we find a significant dependence of absorption spectra on the strength of diphenylaminophenyl donors.

I. Introduction

Organic chromophores possessing a high degree of π-conjugation are ideal materials for advanced electronic and photonic applications including optical information processing, photovoltaic cells, photodynamic therapy agents, and many other applications.[1] These properties are due to the stable, macrocyclic, conjugated network of π-electrons, which lead to high electrical polarizabilities and rapid nonlinear optical responses (NLO) of the charge density to the applied intense electromagnetic fields. The optical spectra and NLO properties of chromophores may be conveniently modified *via* a change of molecular architecture, substituent groups, and substitution patterns.[2−7]

The need to understand the relationship between chemical structure and NLO properties at the molecular level has led to the exploration of various electron-donor and electron-acceptor end-groups linked through π-bridges of various natures and lengths. One of the most successful NLO systems combines the 2-dicyanomethyl-3-cyano-4,5,5-trimethyl-2,5-dihydrofuran acceptor end-group, often referred to as "tricyanofuran" (TCF), with substituted amine donor end-groups. The π-linkage may include thienylvinylene (FTC)[3] and tetraene (CLD)[4] type bridges.

These push−pull chromophores exhibit superior NLO properties, characterized by a large hyperpolarizability (β) values and good thermal and chemical stabilities. Devising efficient electro-optic (EO) materials that also have good

* Corresponding author e-mail: jandzelm@arl.army.mil.
† U.S. Army Research Laboratory, Aberdeen Proving Ground.
‡ The Hebrew University of Jerusalem.
§ Pacific Northwest National Laboratory.

optical transparency represents a considerable development dilemma. For instance, the transparent chromophores in a visible yellow spectral range are typically small molecules with a low EO response, while molecules possessing large β-values are frequently opaque or, at best, have small windows of visibility. The visibility window in these NLO frameworks is generally bracketed by two kinds of electronic transitions.[2−7] The blue-shifted transition is a local excitation of $\pi-\pi^*$ character, while the red-shifted transition is characterized by a significant intramolecular charge-transfer (CT). The CT peak is typically the maximum absorption peak (λ_{max}), while the $\pi-\pi^*$ transition has a much lower oscillator strength and could be denoted as λ_{max-1}. An increase in the β-value due to a change in end-groups is usually accompanied by a red-shift of the λ_{max} peak. The optimal chromophores, which are transparent in the visible spectrum, would have a red-shifted λ_{max}-peak, while the λ_{max-1}-peak should be blue-shifted to open a large visibility window. In this work, we assess the accuracy of computational methods in predicting both the λ_{max}- and λ_{max-1}-peaks of optical absorption spectra. A reliable computational method, if available, would allow the prediction of both the position and intensity of either peak with a small, consistent error. This would facilitate the formulation of a relationship between molecular structure and the UV−vis absorption spectra of chromophores and contribute to the design of chromophores which are transparent in the visible.

Due to the almost limitless optimization space of chromophore structures, computational tools are increasingly being used to assist experimental efforts in designing optimal chromophores. By far the most often used computational approaches are semiempirical and DFT methods. The semiempirical computational methods, such as INDO/CIS, are fast and accurate for chromophores similar to systems for which the method was parametrized.[8]

DFT offers the best compromise between accuracy and computational performance for typical chromophores of about 100 atoms. The electronic spectra are often calculated using time-dependent density-functional theory (TD-DFT)[9,10] and numerous successes of this technique have been recently reviewed.[11]

Recent papers on visible absorption spectra of numerous organic dyes[12] found that the DFT method at the PBE0 level delivers overall the best estimates of λ_{max} values. In the case of push−pull π-conjugated chromophores, it is well-known that the conventional exchange-correlation potentials of DFT fail to predict accurately optical properties.[13] Unacceptably large errors were reported for chromophores with increasing chain lengths of the conjugated framework.[14,15] The failure of TD-DFT calculations has been associated with the incorrect asymptotic behavior of typical exchange-correlation (XC) potentials. These approximate XC potentials do not correctly treat self-repulsion, which leads to the so-called self-interaction error (SIE).[11] As a result, the excited states of dyes, in particular the charge-transfer (CT) of chromophores, are poorly described.[13,16,17] Recently, a new class of DFT functionals has been proposed, which includes a growing fraction of "exact" exchange as the distance increases.[18−24] These long-range corrected (LC) functionals

have been found to provide improved correlation between calculated and experimental optical properties of several dyes.[12,24] The recently proposed BNL functional,[21,23] which falls into this LC-class, is of particular interest for studying push−pull dyes because it has been shown to reproduce exactly the CT excitation for intermolecular complexes.[23] The full long-range cancelation of self-repulsion in BNL restores the long-range effective potential 1/r and also allows an improved CT excitation description for intermolecular complexes. In this paper, we report the implementation of the BNL method in the parallel NWChem program package.[25] The performance of the BNL method will be compared with other DFT approaches, in particular with the CAM-B3LYP functional.[12,22]

The focus of this research is push−pull chromophores with large β-values for which accurate experimental UV−vis spectra are available. We present representative results for four types of molecules: 2-dicyanomethylen-3-cyano-4-{2-[E-(4-N,N-di(2-acetoxyethyl)-amino)-phenylene-(3,4-dibutyl)-thien-5]-E-vinyl}-5,5-dimethyl-2,5-dihydrofuran[3,5] (denoted FTC), CLD,[4,5] recently synthesized, highly efficient chromophores[6] based on the (4-diaryloamino) phenyl electron donors and 2-dicyanomeththylen-3-cyano-4-methyl-5-phenyl-5-trifluoro-methyl-2,5-dihydrofuran (CF3-TCF) electron acceptors, and the TCV chromophore with a strong acceptor (tricyanovinyl) and donor (dimethylamine).[26] This selection allows us to evaluate and validate the theoretical methods for calculating absorption spectra of this important class of molecules.

Several studies have been published critically comparing the performance of various computational methods for chromophores with TCF acceptors. The most comprehensive comparison of Hartree−Fock (HF), INDO, and DFT methods in calculations of dipole moments, polarizabilities, and β-values was published by Isborn et al.[27] They found that the relative merits of molecules are consistently predictable with every method. However, no optical absorption spectra were evaluated in this work. The effect of conformation and isomerization on NLO properties of the FTC chromophore have been reported by Kinnibrugh et al.[28] The β-values change significantly even for nearly degenerate conformers. However, the effect of isomerization on the absorption spectra of FTC was not investigated. The INDO method was recently used to compare properties of several chromophores with strong nitrile-based acceptor end-groups and various modifications of the π-bridge.[29,30] The trends in β- and λ_{max}-values have been correlated with modifications of the chemical structure. The errors in absolute values of λ_{max} as calculated with the INDO approximation can reach 150 nm or more, which is almost as large as the entire visibility window of interest. Moreover, no λ_{max-1}-peak was calculated, thus making it difficult to evaluate the applicability of the INDO method in predicting the transparency window of such chromophores.

This paper is organized as follows. First, we will describe the implementation of the BNL method in the NWChem program and verify the sensitivity of the method to the choice of technical parameters, such as basis set or model of solvation. We will then verify the performance of standard

Optical Spectra of TCF-Chromophores

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2837**

DFT functionals for chromophores with the TCF motif and study the dependence of absorption spectra on the choice of low lying conformers of FTC and CLD. This will allow us to select the most stable conformer for a detailed study of comparing the performance of the DFT-LC method for several chromophores with TCF-type acceptors. We will judge the applicability of DFT-LC functionals in predicting the visibility window by comparing positions of both the λ_{max}- and λ_{max-1}-absorption peaks and their corresponding oscillator strengths to experimental data. Finally, the BNL method will be applied to study the effect of varying substituents in models of CF3-TCF-type chromophores on the respective optical spectra. The optimized structures are provided as Supporting Information.

II. Computational Methodology

The structures of all molecular systems were calculated using the Gaussian03 program.[31] The B3LYP exchange-correlation functional and the 6-31(d,p) basis set were used in the gas phase calculations.[31-34] The geometry optimization in a solvent was done using the PCM model as implemented in Gaussian03.[35] The vertical excitation energies were calculated using the time-dependent DFT (TDDFT)[10,36] module implemented in the NWChem computational chemistry package[25] developed at the Pacific Northwest National Laboratory. This program allows us to compare the performance of several long-range corrected (LC) DFT approaches[21-23] as well as other typical DFT exchange-correlation functionals, which is one of the main goals of this paper. To simulate the effect of the solvent on the excitation energies, we have used the COSMO solvation model[37] implemented in NWChem.

Three classes of DFT functionals are investigated here. The first class consists of the fundamental local density (LDA) and generalized-gradient (GGA) approximations[38] represented by SVWN and PBE,[39] respectively. LDA and GGA depend on the electron density and its gradient, respectively.[38] The second class is comprised of three hybrid methods, B3LYP,[32-34] PBE0,[40] and BH&H.[41] In all these hybrids, the percentage of Hartree−Fock exchange is constant at each point in space, and it amounts to 20, 25, and 50% of HF exchange for B3LYP, PBE0, and BH&H, respectively. The hybrid functionals yield quite accurate transition energies for many organic dyes; however, they poorly describe charge-transfer states that dominate the λ_{max}-transitions of the molecules studied in this paper. The most successful strategy used to solve this problem involves the use of the LC DFT functionals.[18-24] These functionals include a growing fraction of Hartree−Fock exchange when the interelectronic distance increases, thus effectively limiting the dominant self-interaction error at large distances. This procedure, originally suggested by Savin,[18] has been further extended resulting in several implementations of LC-DFT functionals. The essence of these functionals lies in partitioning the Coulomb operator into long- and short-range components[22,24] in an Ewald-like fashion as

$$\frac{1}{r} = \frac{1}{r}\{1 - [\alpha + \beta \mathrm{erf}(\gamma r)]\} + \frac{1}{r}[\alpha + \beta \mathrm{erf}(\gamma r)] \tag{1}$$

where γ is the attenuation or range-separation parameter, while α and β define the contributions of the Hartree−Fock exchange. With this, the exchange energy, E_x, can be partitioned into short- and long-range components, respectively

$$E_x = E_x(short) + E_x(long) \tag{2}$$

The short-range part of the exchange is treated with traditional DFT, while the long-range component is treated with "exact" exchange. A crucial issue in the construction of these functionals is the choice of parameters α, β, and γ. The parameters α and β satisfy the relations $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, and $0 \leq \alpha + \beta \leq 1$.[20-24] In this work, we have used the CAM-B3LYP functional of Handy and co-workers, which applies the Coulomb attenuation method to B3LYP.[22] We have used two parametrizations of CAM-B3LYP, the original one with $\alpha = 0.19$, $\beta = 0.46$ and a modified one with $\alpha = 0.19$ and $\beta = 0.81$.[22] The first parametrization yields $\alpha + \beta = 0.65$, so no exact asymptote of the exchange potential is reached. The second parametrization yields $\alpha + \beta = 1$ indicating that the $E_x(long)$ term, asymptotically (at $r \rightarrow \infty$) is equal to Hartree−Fock exchange.

In the BNL LC approach[21,23] the electron repulsion is also separated into long- and short-range parts according to eq 1; however, the parameters are set to $\alpha = 0$ and $\beta = 1$. Therefore, the short-range part is treated with traditional DFT, while the long-range component, unlike the CAM-B3LYP functional described above, attains the "exact" exchange value only in the asymptotic limit. In the BNL functional, the exchange-correlation energy per particle is separated into γ-dependent exchange and correlation parts respectively as follows

$$\varepsilon_{xc}^{\gamma}(n) = \varepsilon_x^{\gamma}(n) + \varepsilon_c^{\gamma}(n) \tag{3}$$

where n represents the density. The exchange energy per particle is presented in Savin's form,[18] where $q = (\gamma)/(k_f)$ and k_f is the local Fermi vector at the given density n. The exchange and correlation energies are represented as

$$\varepsilon_c^{\gamma}(n, |\nabla n|) = \varepsilon_c^{LYP}(n, |\nabla n|) - w\varepsilon_x^{\gamma}(n) \tag{4}$$

where the major part of the correlation energy has been represented with the GGA functional LYP.[32] It has been found that subtracting a small piece of the Savin exchange (eq 4) with a small mixing factor w improves the overall performance of the functional.[23] It can be shown that the theory becomes more HF-like for large γ and behaves more local DFT-like for small γ. A crucial issue in the construction of LC functionals is the choice of γ. It has been shown[23] that this parameter can be optimized using an extensive molecular training set, but recent calculations[42] have revealed that this parameter is in fact system-dependent. The value $\gamma = 0.5$ was used in the original paper, while γ as small as $\sim$0.1 has been recommended for some metallic systems.[42] In this study, we performed further validation of this parameter on the molecular systems considered in this paper.
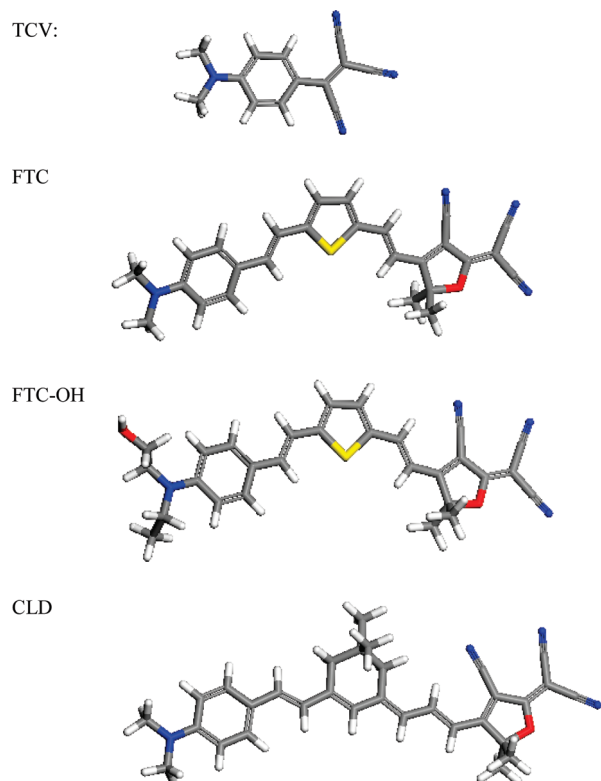
TCV:

FTC

FTC-OH

CLD

Figure 1. The B3LYP/6-31G** optimized structures of TCV, FTC, FTC—OH, and CLD chromophores.

To estimate the vertical excitation energies, we have used the TD-DFT approach.[9,10] TD-DFT is an extension of DFT in which electronic excitations are calculated as the poles of the electron density response.[9,10] We also tested the Tamm-Dancoff approximation,[43] as it is computationally more efficient than solving the complete TD-DFT equations. We found a systematic blue-shift of spectra and decided not to use this approximation. We have frozen the atom-like core orbitals in TD-DFT calculations.

Since the long-range part of LC-functionals has to be calculated explicitly, the 2-electron integrals have to be considered with care. In particular, the exchange interactions have to be treated separately from the pure Coulomb interactions because the attenuation just affects the exchange. We have implemented two approaches to deal with this. The first approach utilizes the well-known Dunlap[44] charge fitting method for the pure Coulomb interactions, whereas the exchange contribution (including the attenuation) is treated in the conventional manner. The second approach involves calculating directly all the exchange integrals. Our charge fitting approach is implemented along the lines of the von Arnim and Ahlrichs implementation.[45] The Coulomb contribution with this approach is evaluated with 3-center integrals. The in-core strategy in NWChem for storing all the 3-center, 2-electron integrals in memory also allows for a much faster evaluation of the Coulomb contribution compared with disk-based or direct approaches. We have used the DFT DGauss Coulomb fitting basis[46] for the charge fitting on all the atoms.

The effect of basis set expansion on UV−vis spectra was studied for the TCV molecule presented in Figure 1. Table 1 shows λ_{max}- and λ_{max-1}-transitions and corresponding

oscillator strengths obtained for basis sets of increasing size. Both hybrid and Coulomb attenuation functionals were tested. In order to compare the BNL and CAM-B3LYP functionals, common parameters, $\gamma = 0.33$ and $\alpha + \beta = 1$, were used for both functionals. For the CAM-B3LYP calculations we have used $\alpha = 0.19$, $\beta = 0.81$ and for the LC-BNL calculations we have used $\alpha = 0$, $\beta = 1$, respectively The extension of basis sets leads to red-shifted spectra by less than 20 nm in the case of the λ_{max}-peak and less than 10 nm for the λ_{max-1}-peak. The oscillator strengths are more sensitive to the basis set choice, particularly for the λ_{max-1}-peak of the CAM-B3LYP method. We also tested the accuracy of fitting functions by comparing results with those obtained using a direct method that does not employ fitting functions. No significant deterioration of results was observed due to the fitting. In the case of very large chromophores, we have used the smaller 6-31G* basis set as it represents the best compromise between computational efficiency and accuracy.

Predicting the effect of a solvent on absorption spectra is a difficult task as, in principle, it requires consideration of both a fast, electronic and a slow, orientational relaxation response of a solvent.[47,48] Solvents of interest have typically very low dielectric constants, and we assume that excitation energies are not too sensitive to the choice of the solvation model. In this work, we have considered the "conductor-like screening model" COSMO as implemented in the NWChem program. In this model, the solute interacts with the solvent represented by a dielectric continuum model.[37,48] The solute molecule is embedded into a cavity surrounded by a dielectric continuum of permittivity ε. The dielectric screening energies for a given geometry scale as $(\varepsilon - 1)/(\varepsilon + x)$ with the dielectric permittivity ε of the screening medium, where x is in the range of $0-2$. The scaling x($= 0$) was used,[49] and it was confirmed that x($= 1/2$) shifts λ_{max} by less than 5 nm, even for the chloroform solvent with a very low dielectric. We have tested several techniques of building the COSMO cavity considering different sets of radii[48,49] and tesselation of the unit sphere as well as octahedral and icosahedral representations of the cavity surface.[25] The absorption spectra of the TCV, FTC, and CLD molecules appear to be little affected, and, therefore, the NWChem default COSMO parameters were used throughout this study. Finally, we have tested the effect of structural changes due to the presence of solvent on the absorption spectra and found a small red shift of less than 5 nm for both peaks. In summary, we expect that the red-shift of absorption spectra due to the choice of basis set or solvent model does not exceed 20 nm. We further assume that the possible discrepancy with respect to experimental values larger than 30 nm is caused by the inaccuracy of the particular DFT functional.

The accuracy of the absorption spectra calculations could be improved by using highly correlated ab initio approaches such as CCSD or CAS-PT2 methods. The CAS-PT2 method could account for the multiconfigurational nature of the chromophore's wave function. It is known that CAS-PT2 underestimates slightly excitation energies and needs an empirical correction.[50] Furthermore, the application of CAS-

Optical Spectra of TCF-Chromophores

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2839**

Table 1. UV-vis Spectra and Oscillator Strength (f) of TCV Calculated at the BNL, B3LYP, and CAM-B3LYP Levels for Various Basis Sets[b]

basis set	BNL		B3LYP		CAM-B3LYP	
	λ_{max} (f)	λ_{max-1} (f)	λ_{max} (f)	λ_{max-1} (f)	λ_{max} (f)	λ_{max-1} (f)
6-31G*	396(0.81)	237(0.21)	451(0.66)	289(0.13)	412(0.78)	255(0.10)
6-31G**	396(0.81)	238(0.21)	452(0.65)	290(0.13)	412(0.77)	255(0.09)
6-311G*	402(0.80)	240(0.19)	456(0.65)	292(0.12)	417(0.77)	257(0.06)
6-311G*[a]	402(0.80)	240(0.20)	456(0.65)	292(0.12)	417(0.77)	258(0.06)
6-31+G*	411(0.83)	243(0.21)	462(0.67)	295(0.12)	423(0.79)	253(0.15)
6-311+G*	412(0.81)	243(0.20)	464(0.66)	295(0.12)	424(0.77)	255(0.16)
6-311++G**	412(0.81)	245(0.15)	464(0.66)	295(0.12)	425(0.77)	256(0.17)

[a] Direct calculations without density fitting functions. [b] The λ_{max}, λ_{max-1} (in nm), and oscillator strengths (f) values are provided. Calculations were done using the COSMO solvation model, assuming the 1,4-dichloroethane solvent.

PT2 becomes prohibitively expensive with sufficiently large active spaces.[12,50] Therefore as a compromise of computational expense and accuracy, we have not performed CAS-PT2 calculations for the TCF-chromophores considered in this paper.

III. Results and Discussion

A. Validation of DFT Methodology. The structures of chromophores with the TCF motif used to validate the performance of DFT functionals are presented in Figure 1. These are the lowest energy structures as calculated at the B3LYP/6-31G(d,p) level of theory in the gas phase. The experimental UV−vis spectra for these molecules in a solvent environment are available and, therefore, the 1,2-dichloroethane solvent was considered for TCV, while cyclohexanone and tetrahydrofuran solvents were used in FTC and CLD chromophore calculations, respectively, for comparison with experiment. The values of the dielectric constants used in the continuous electrostatic models were 10.3, 18.2, and 7.6, respectively. The spectra for FTC and CLD chromophores were measured with a hydroxyl-diethylamine donor instead of the simplified dimethylamine shown in Figure 1.[5] We have performed calculations with the diethyl donor and its hydroxyl derivative (FTC−OH chromophore displayed in Figure 1) finding that the main features of the spectra, such as absorption peak (λ_{max}), were shifted by less than 10 nm with the increased size of a donor group.

Table 2 includes TD-DFT results of vertical excitation spectra including λ_{max}, λ_{max-1}, and corresponding oscillator strengths at various levels of DFT theory. The Hartree−Fock, hybrid DFT functionals (BHLYP, PBE0, B3LYP), gradient corrected DFT (PBE), and the simplest, density-dependent, SVWN functionals were used. The experimental oscillator strengths were obtained by fitting two Gaussians to the two major peaks in the experimental spectrum. The ratio of the area under each Gaussian was used to estimate the respective relative oscillator strengths.[51] This is a crude approximation of the transition strength; nevertheless, it can shed light on the relative importance of transitions and width of the visibility window.

As expected, the accuracy of results in matching the experimental data depends strongly on the method used. The PBE and SVWN methods reproduce within 30 nm major absorption peaks but quite poorly the relative oscillator strengths of the TCV chromophore. However, this level of theory fails completely for the larger FTC and CLD systems.

Table 2. Absorption Spectra (λ_{max} and λ_{max-1} in nm) and Oscillator Strengths (f) of TCV, FTC, and CLD for Various DFT Functionals[a]

functional	λ_{max}	f_{max}	λ_{max-1}	f_{max-1}	rel. f[b]
		TCV			
PBE	504	0.55	328	0.14	0.25
PBE0	451	0.70	285	0.12	0.17
BHLYP	411	0.82	254	0.12	0.14
B3LYP	462	0.67	294	0.12	0.18
SVWN	509	0.55	330	0.14	0.25
HF	355	0.96	222	0.16	0.17
expt	518		292		0.08
		FTC			
PBE	766	1.07	541	0.83	0.78
PBE0	637	1.54	439	0.49	0.32
BHLYP	514	2.40	350	0.23	0.10
B3LYP	663	1.43	458	0.57	0.40
SVWN	785	1.01	544	0.86	0.85
HF	458	1.90	295	0.14	0.07
expt	650		400		0.26
		CLD			
PBE	728	1.45	555	0.91	0.63
PBE0	621	2.18	450	0.39	0.18
BHLYP	548	2.55	372	0.05	0.02
B3LYP	639	2.07	468	0.47	0.23
SVWN	735	1.44	559	0.92	0.64
HF	462	2.75	238	0.20	0.07
expt	660		420		0.04

[a] COSMO/6-31+G*//B3LYP/6-31G**. [b] The relative value of oscillator strengths $f_{\lambda max}/f_{\lambda max-1}$.

The hybrid functionals, which yield erroneously blue-shifted absorption peaks in the case of TCV, perform much better for the FTC and CLD chromophores. The accuracy improves with decreasing amount of Hartree−Fock exchange in the hybrid functional. The B3LYP method shows improved values for λ_{max}, while the values for λ_{max-1} are significantly red-shifted. The PBE0 method could be recommended for FTC and CLD systems as the best compromise. It both underestimates the λ_{max} peak and overestimates λ_{max-1} by about 40 nm. The relative values of the oscillator strengths are predicted to be too high particularly for the CLD chromophore. The use of PBE0 was recently recommended for calculations of the $\pi-\pi^*$ transitions for numerous organic dyes.[12b]

We now turn to evaluate the performance of Coulomb attenuation functionals, CAM-B3LYP and BNL. Tables 3 and 4 display absorption spectra and oscillator strengths of TCV, FTC, and CLD as a function of the attenuation parameter, γ. Regardless of the type of LC-DFT functional,

Table 3. UV-vis Spectra and Oscillator Strength (f) of TCV, FTC, and CLD Chromophores (Figure 1) Calculated at the CAM-B3LYP Level with Various Attenuation Parameters, γ^a

	TCV		FTC		CLD	
method (γ)	λ_{max} (f)	λ_{max-1} (f)	λ_{max} (f)	λ_{max-1} (f)	λ_{max} (f)	λ_{max-1} (f)
experiment[b]	518(1.00)	292(0.08)	650(1.00)	400(0.26)	660(1.00)	420(0.04)
γ^c						
0.01	464(0.67)	295(0.12)	657(1.61)	456(0.81)	642(2.04)	472(0.49)
0.05	462(0.67)	293(0.12)	632(1.79)	437(0.63)	627(2.20)	452(0.32)
0.10	453(0.70)	283(0.13)	583(2.08)	392(0.29)	597(2.36)	410(0.10)
0.15	438(0.74)	267(0.16)	547(2.21)	333(0.19)	572(2.46)	317(0.14)
0.20	424(0.78)	256(0.20)	522(2.25)	313(0.21)	553(2.45)	300(0.15)
0.33	399(0.85)	235(0.24)	485(2.29)	280(0.20)	519(2.48)	274(0.14)
γ^d						
0.01	464(0.67)	295(0.12)	657(1.61)	456(0.81)	642(2.04)	472(0.49)
0.05	462(0.67)	293(0.12)	642(1.71)	445(0.70)	627(2.19)	452(0.32)
0.10	457(0.68)	288(0.11)	609(1.92)	418(0.49)	614(2.26)	434(0.23)
0.15	449(0.71)	280(0.09)	584(2.04)	395(0.33)	596(2.35)	413(0.13)
0.20	440(0.74)	266(0.10)	565(2.11)	378(0.23)	582(2.41)	327(0.13)
0.33	423(0.79)	253(0.14)	528(2.21)	320(0.19)		

^{a} λ_{max} and λ_{max-1} (nm) with significant values of oscillator strength are reported. The most stable conformer was used within the COSMO environment. The 6-31+G* basis set was used. ^{b} Relative values of oscillator strengths are reported. ^{c} CAM-B3LYP parametrization α/β with $\alpha = 0.19$, $\beta = 0.81$ is used. ^{d} CAM-B3LYP parametrization α/β with $\alpha = 0.19$, $\beta = 0.46$ is used.

Table 4. UV-vis Spectra and Oscillator Strength of TCV, FTC, and CLD Chromophores (Figure 1) Calculated at the BNL Level with Various Attenuation Parameters, γ^a

	TCV		FTC		CLD	
method (γ)	λ_{max} (f)	λ_{max-1} (f)	λ_{max} (f)	λ_{max-1} (f)	λ_{max} (f)	λ_{max-1} (f)
experiment[b]	518(1.00)	292(0.08)	650(1.00)	400(0.26)	660(1.00)	420(0.04)
0.01	510(0.54)	330(0.13)	781(0.97)	536(1.04)	732(1.47)	561(0.83)
0.05	507(0.55)	326(0.12)	722(1.25)	511(0.84)	700(1.79)	528(0.58)
0.10	492(0.59)	309(0.13)	641(1.68)	444(0.48)	649(2.20)	462(0.21)
0.15	469(0.65)	291(0.14)	584(1.89)	396(0.26)	608(2.35)	416(0.06)
0.20	448(0.71)	272(0.14)	546(1.98)	365(0.22)	579(2.39)	386(0.01)
0.33	411(0.83)	243(0.21)	490(2.25)	293(0.24)	531(2.41)	345(0.04)

^{a} Structure, basis set, and solvation model are the same as in Table 3. ^{b} Relative values of oscillator strengths are reported.

we observe a red-shift of the spectra with decreasing γ. This is consistent with the results obtained using standard DFT functionals, as reported in Table 2. Decrease in γ diminishes the role of Hartree−Fock exchange, which accounts for a significant red-shift of spectra for DFT functionals such as PBE. For the same value of γ, the spectra are more red-shifted for BNL compared to the CAM-B3LYP functional. For BNL, one can identify the common value of γ approximately equal to 0.1, that results in a λ_{max} close, within 20 nm, to the experimental values for all test molecules. In the case of the FTC molecule, both CAM-B3LYP and BNL perform reasonably well, if γ close to 0.05 and 0.1 is used, respectively. The positions of the λ_{max}-peaks, and the relative strengths of peaks are well reproduced; however, the λ_{max-1}-peak is red-shifted by ca. 40 nm. The sensitivity to γ-values of absorption spectra for various systems was already reported in the literature.[42,52−55] The results of Table 3 and 4 suggest also a strategy of using different values of γ to calculate various excitations. In the case of BNL, the optimal position of the λ_{max}-peak can be obtained using $\gamma = 0.1$, while the larger value of $\gamma = 0.15$ may be needed to improve the prediction for the λ_{max-1}-peak. This would however worsen the description of the relative strength of both peaks for the FTC molecule while improving it for the CLD chromophore.

Recently the performance of DFT functionals was analyzed, and a diagnostic test was proposed to predict the accuracy of excitation energies.[56] This test was based on the spatial overlap of occupied and virtual orbitals, Λ. One of the molecules studied in that paper, DMABN, is similar to the TCV molecule studied by us, the difference being a cyano acceptor vs the TCF acceptor, respectively. In the case of the PBE functional, the DMABN molecule exhibits a large value of Λ and also quite an accurate description of the CT excited state.[56] We did not implement the diagnostic test in this paper; however, following conclusions by Peach et al.[56] we can predict that less accurate results of excitation energies for FTC and CLD molecules indicate a small Λ value, while excellent results for TCV molecule are likely related to a large Λ (see Table 2).

The absorption spectra results could perhaps be further improved by considering statistical corrections that were recently applied for the LC-functionals in calculations of $\pi−\pi^{*[12b]}$ and n-π^* excitations.[57]

To analyze the nature of the discussed λ_{max} and λ_{max-1} states, the transition orbitals are frequently displayed, and their contributions to the excitation energies as found from the TD-DFT calculations are discussed.[13,14,28] As expected, the excitation at λ_{max} has a charge-transfer character as it is dominated by transitions from HOMO to LUMO orbitals of the chromophore. The HOMO orbital is predominantly localized on a donor site of the chromophore, while the LUMO orbital occupies the acceptor site. The HOMO→LUMO transition gives rise to a well-known intramolecular charge-

Optical Spectra of TCF-Chromophores

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2841**

Table 5. Analysis of the Electronic Excitations (CT and $\pi-\pi^*$ Transitions) for the FTC Chromophore Calculated at the BNL Level with Various Attenuation Parameters, γ^a

method (γ)	CT		$\gamma-\gamma^*$		H-L[b]
	H→L	H-1→L	H-1→L	H→L+1	
0.01	0.96	0.24	0.77	0.54	1.33
0.05	0.99	0.13	0.82	0.53	2.36
0.10	0.99	0.10	0.89	0.37	3.26
0.15	0.95	0.25	0.91	0.08	3.96
0.20	0.90	0.34	0.86	0.25	4.45
0.33	0.84	0.44	0.68	0.57	5.34

[a] The absolute values of weights for two major orbital transitions of each excitation are shown. Computational parameters are the same as in Table 4. [b] The HOMO(H) - LUMO(L) gap in eV.

Figure 2. Definition of dihedral angles used to study conformations of FTC and CLD chromophores.

transfer state.[11,13,14] The excitation at λ_{max-1} can be recognized as a $\pi-\pi^*$ transition, as it involves orbitals below the HOMO and above the LUMO that span the entire framework of the chromophore. We have analyzed orbital contributions to the electronic transitions for the FTC chromophore calculated using the BNL approach for various attenuation parameters, γ. The absolute values of weights and the character of major orbital transitions are presented in Table 5. This confirms that the λ_{max}-peak can be interpreted as a charge-transfer peak because the major contribution comes from the HOMO to LUMO excitation. Similarly the λ_{max-1}-peak involves π-electron excitations, and it is dominated by the HOMO-1 to LUMO excitation. It is interesting to note that the λ_{max}-peak has increased charge-transfer character around $\gamma = 0.1$, as the contribution of the HOMO-1→LUMO transition shows a minimum at this value. Similarly the λ_{max-1}-peak has the strongest $\pi-\pi^*$ character about $\gamma = 0.15$. This strengthening of the charge-transfer and $\pi-\pi^*$ characters coincides with an improved performance of the BNL functional in predicting λ_{max}- and λ_{max-1}-peaks, respectively. As is evident from Table 4, the excitation energies are consistently red-shifted with decrease of the attenuation parameter. This corresponds to closing the HOMO−LUMO gap as shown in Table 5.

In summary, assuming a target accuracy of 30 nm in the prediction of the most important λ_{max}-transition, we can recommend the BNL method with attenuation parameter $\gamma = 0.1$ for systems studied in this work. For the FTC and CLD chromophores, the hybrid functionals B3LYP and PBE0 as well as CAM-B3LYP with $\gamma = 0.05$ would also yield the λ_{max}-transition within our target accuracy.

B. Absorption Spectra of FTC and CLD Conformers. Extended molecules such as FTC and CLD have several rotatable single bonds, which lead to the existence of conformers with low relative energies. The manifold of rotational isomers of the FTC chromophore was recently studied by Kinnibrugh et al.[29] The authors found relatively low barriers for rotations along the single bonds, indicating that the lifetime of a given rotamer is on the order of 2.5 ms, thus making it quite plausible that several rotamers coexist in the gas phase as well as in a low-viscosity solution.

The eight low-energy conformers of each molecule were optimized, and a significant dependence of NLO properties on the conformer's structure was observed.[29] The values of the hyperpolarizability projected on the axis of the dipole moment can va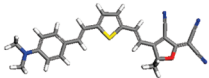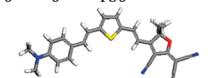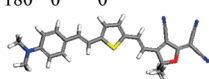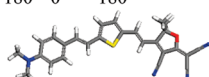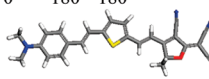ry by a factor of 2 between close-lying conformers. The effect of the conformational space on the UV−vis spectra has not been considered until now.
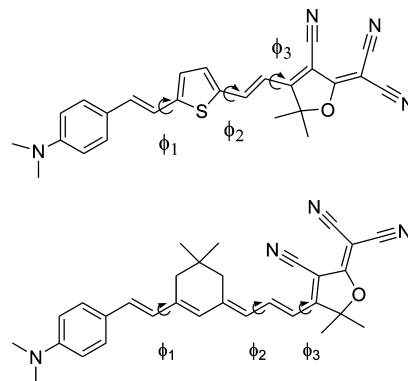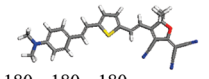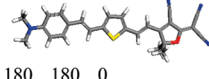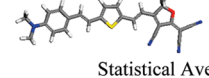
We have optimized the structures of the eight conformers that can be created by rotations around the single bonds as presented in Figure 2. The structures of the conformers were optimized at the B3LYP/6-31G** level of theory in the gas phase (G), and the absorption spectra were calculated in the presence of solvent (S) modeled at the COSMO level of theory. Tables 6 and 7 present results of absorption spectra for FTC and CLD conformers, respectively. The energy differences between conformers, in kcal/mol, are presented in the gas phase (G) as well as in the solvent, S. Both the G and S calculations utilize the gas-phase-optimized structure. The effect of solvent can change the ordering of conformers separated by less than 2 kcal/mol for both the FTC and CLD chromophores. The structures of the gas-phase chromophores are close in energy, within 3 and 8 kcal/mol for FTC and CLD systems, respectively. The presence of solvent changes these values by 1 kcal/mol at most. Since the solvent field is much stronger in the case of FTC than CLD, we have optimized the structures of FTC at the B3LYP/6-31G** level with the PCM solvation model. The inclusion of the PCM solvation model in the geometry optimization changes the order of the conformers and overall spreads the distribution of the eight FTC isomers to 5.7 kcal/mol. This solvent-optimized geometry however has only a minor effect on the absorption spectra, and while the statistical average of λ_{max-1} appears to be red-shifted by about 11 kcal/mol, the average λ_{max} value is virtually unchanged at 652 nm. We can hence adopt a computationally much simpler strategy of using the gas-phase-optimized geometry to calculate the absorption spectra in the solvent environment of the COSMO model.

Results in Tables 6 and 7 indicate that the Boltzmann-weighted root-mean-square deviation over all conformers was less than 10 nm for λ_{max} and less than 5 nm for λ_{max-1}. While the addition of solvation effects shifts the spectra appropriately, this does not change the influence of the conformation, neither at the gas phase nor at the solvated geometry. It is therefore acceptable to use the lowest-energy geometry in the gas phase with solvation effects for the computation of spectra. The results obtained for the lowest-energy conformer are within 12 nm or better of the statistical average of all conformers. Equally important is the analysis of the relative strength of the transitions as measured by the

Table 6. Optimized Gas-Phase Geometries and Optical Spectra for FTC[a,b,c,d]

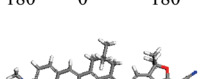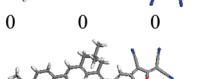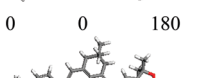| ϕ_1	ϕ_2	ϕ_3	ΔE^b	ΔE^c	λ_{max}	f_{max}	λ_{max-1}	f_{max-1}
0	0	0	0.0	0.44	661	1.45	457	0.55
0	0	180	0.41	0.55	659	1.35	454	0.55
180	0	0	0.86	3.77	625	1.71	449	0.50
180	0	180	1.43	0	644	1.55	452	0.67
0	180	180	1.74	1.82	670	1.55	456	0.70
0	180	0	1.99	1.21	668	1.47	461	0.51
180	180	180	2.55	2.51	652	1.63	452	0.79
180	180	0	2.79	1.18	653	1.55	456	0.66
Statistical Average					649	1.49	467	0.62
Statistical Deviation					9	0.08	2.9	0.06
Relative f						1.00		0.41±0.06
Experiment					650	1.00	400	0.26

[a] λ_{max}, λ_{max-1}, and corresponding oscillator strengths (f) from B3LYP/COSMO/6-31+G*//B3LYP/6-31G**. [b] Energy differences from B3LYP/6-31G**//B3LYP/6-31G**. [c] Energy differences from B3LYP/COSMO/6-31G**//B3LYP/6-31G**. [d] Dihedral angles are defined in Figure 2.

Table 7. Optimized Gas-Phase Geometries and Optical Spectra for CLD[a,b,c,d]

ϕ_1	ϕ_2	ϕ_3	ΔE^b	ΔE^c	λ_{max}	f_{max}	λ_{max-1}	f_{max-1}
180	180	180	0.0	0.27	639	2.07	468	0.47
180	180	0	1.11	0	649	1.95	463	0.83
0	180	180	2.46	3.25	640	1.92	473	0.40
0	180	0	3.50	2.27	640	1.90	469	0.58
180	0	0	4.84	6.01	653	1.92	469	0.52
180	0	180	5.91	5.54	650	1.97	470	0.52
0	0	0	7.34	6.93	650	1.86	473	0.44
0	0	180	8.31	6.55	645	1.92	474	0.37
Statistical Average					646	1.94	470	0.52
Statistical Deviation.					6	0.06	4	0.14
Relative f						1.0		0.26±0.08
Experiment					660		420	

[a] λ_{max}, λ_{max-1}, and corresponding oscillator strengths (f) from B3LYP/COSMO/6-31+G*//B3LYP/6-31G**. [b] Energy differences from B3LYP/6-31G**//B3LYP/6-31G**. [c] Energy differences from B3LYP/COSMO/6-31G**//B3LYP/6-31G**. [d] Dihedral angles are defined in Figure 2.

ratio of oscillator strengths. In the case of the lowest-energy conformers this ratio is within 10% and 15% of the statistical average for the FTC and CLD chromophores, respectively.

Overall, the B3LYP functional performs very well in reproducing the position of the charge-transfer λ_{max}-peak, while yielding the λ_{max-1}-peak significantly red-shifted by about 50 nm, that is outside of our error expectation of about 30 nm. The utility of the B3LYP functional is further diminished considering that the intensity of the $\pi-\pi^*$ peak appears to be erroneously much stronger than needed. Thus the visibility window, as predicted by the B3LYP method, is much narrower than observed in experiment. The conclusion of this section is transferable to other DFT calculations including the BNL functional. Kinnibrugh et al.[29] recently have found that the distribution of values of hyperpolarizability tensors for FTC conformers is fairly similar regardless of the B3LYP or PBE method used. Therefore, we expect that the optical properties of the lowest energy conformer are representative of the entire ensemble of chromophore conformers, unless external fields, such as an electric field, are imposed.

C. Absorption Spectra of Diarylaminophenyl-Containing Chromophores. The highly efficient and thermally stable NLO chromophores based on (4-diaryloamino) phenyl

donors have been recently synthesized and the absorption spectra have been published,[6] thus providing an opportunity to validate further the BNL technique. These chromophores contain a very strong electron-withdrawing acceptor, 2-dicyanomethylen-3-cyano-4-methyl-5-phenyl-5-trifluoro-methyl-2,5-dihydrofuran (CF3-TCF), linked to the donor *via* a π-conjugated bridge containing thiophene or isophore rings. In this work, we evaluate the performance of DFT in predicting spectra of thiophene-based chromophores (A1-A4) as presented in Figure 3 using the protocol derived in the previous sections. The geometry was optimized in the gas phase at the B3LYP/6-31G** level of theory and was subsequently used in TD-DFT calculations at the BNL/COSMO/6-31G* level of theory. A solvent with a dielectric constant of 4.9 ε_0 corresponding to chloroform was used in the COSMO simulations. Following the recommendation of section A, the three best DFT functionals (BNL, B3LYP, and PBE0) for the FTC-type chromophores were used. The calculated vertical excitation energies are presented in Table 8. The BNL functional was applied with the attenuation parameter set to 0.1. Inspection of Table 8 confirms that both
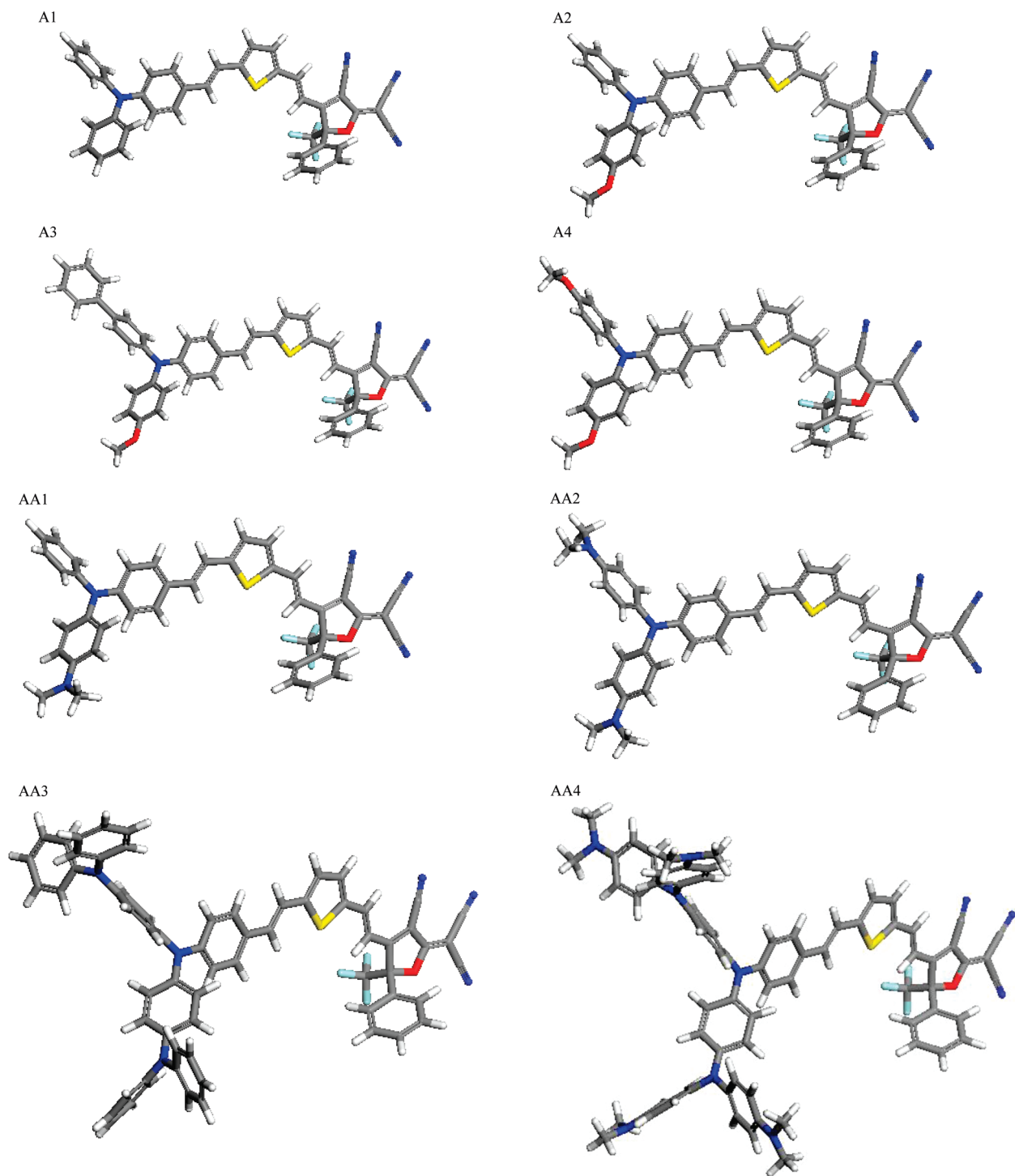
Optical Spectra of TCF-Chromophores

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2843**

Figure 3. The B3LYP/6-31G** optimized structure of A1-A4 and AA1-AA4 chromophores.

PBE0 and BNL predict the positions of excitation levels, exceptionally well. Unlike the parent FTC molecule, the λ_{max-1}-excitations are also very well reproduced. Analysis of the relative intensity of the two main excitations reveals that BNL is superior. This has important consequences in the selection of an appropriate method to predict reliably the visibility window of chromophores.

The λ_{max}- and λ_{max-1}-transitions can be identified as a charge-transfer- and a $\pi-\pi^*$-transition, since the major contributions are from HOMO→LUMO and HOMO-1→LUMO excitations, respectively. A systematic red-shift of transitions can be seen in Table 8 that correlates well with

increasing donor strength. With two additional methoxy groups on the donor, λ_{max} of A4 is shifted by about 50 nm compared to the A1 structure, which exhibits λ_{max} at 685 nm. This shift as calculated at the BNL level matches very well experimental observation. The increased charge transfer to the π-conjugated bridge of the chromophore can also be correlated with the increased dipole moment from 25.3 to 30.7 D for the A1 and A4 chromophores, respectively. We have optimized further the donating strength of aminophenyl donors and replaced the methoxy groups by methyloamino groups resulting in structures AA1-AA4 as shown in Figure 3. The UV−vis spectra for the three highest excitations with

Table 8. UV-vis Spectra and Oscillator Strength of A1-A4 and Chromophores (Figure 3) Calculated at the B3LYP, PBE0, and BNL Levels and Compared with Experimental Results from Ref 6

	λ_{max} (f_{max})	λ_{max-1} (f_{max-1})	f_{max-1}/f_{max}
		A1	
B3LYP	729(1.22)	500(0.88)	0.72
PBE0	687(1.39)	479(0.76)	0.55
BNL	685(1.53)	475(0.50)	0.32
expt	694	485	0.06
		A2	
B3LYP	750(1.21)	514(0.92)	0.76
PBE0	708(1.34)	492(0.82)	0.61
BNL	706(1.51)	486(0.57)	0.38
expt	717	500	0.20
		A3	
B3LYP	762(1.20)	519(1.02)	0.85
PBE0	716(1.35)	497(0.91)	0.67
BNL	708(1.59)	489(0.61)	0.38
expt	721	500	0.20
		A4	
B3LYP	782(1.19)	525(0.98)	0.82
PBE0	737(1.32)	504(0.89)	0.67
BNL	735(1.49)	497(0.64)	0.42
expt	745	500	0.35

Table 9. UV-vis Spectra, Oscillator Strength, and Dipole Moments (μ) of AA1-AA4, Chromophores (Figure 3) and Analysis of Electronic Excitations As Calculated at the BNL Level

chromophore	λ_{max} (f_{max})[a]	λ_{max-1} (f_{max-1})	λ_{max-2} (f_{max-2})	μ
AA1	763(1.22)	545(0.70)[b]	460(0.23)[c]	29.3
AA2	842(1.24)	547(0.79)[c]	544(0.12)[b]	33.9
AA3	776(1.29)	544(0.80)[c]	453(0.22)[d]	27.4
AA4	957(0.88)	640(1.15)[c]	498(0.32)[d]	36.2

[a] Major contribution from the HOMO(H)→LUMO(L) transition. [b] H-1→L. [c] H-2→L. [d] H-3→L, H→L+1.

significant oscillator strengths are presented in Table 9. We observe a significant red-shift of the charge-transfer peak that correlates fairly well with the increased dipole moment of the chromophore. The λ_{max}-peak arises from HOMO→ LUMO transitions and therefore can be interpreted as a charge-transfer transition. In the case of AA4, this is no longer the strongest excitation; nevertheless, we maintain the notation for consistency. λ_{max-1} is due to a $\pi-\pi^*$-transition, as it arises from HOMO-1→LUMO excitations for AA1 and HOMO-2→LUMO for other chromophores. λ_{max-2} for AA3 and AA4 arises from orbitals below the HOMO level and also involves higher virtual orbitals. The $\pi-\pi^*$-excitation is the strongest peak for the AA4 chromophore.

In conclusion, by adjusting the donor strength, we succeeded in pushing the charge-transfer peak to the far-red. However, excitations of $\pi-\pi^*$-character are strengthened and even further red-shifted. Therefore, it is unlikely that the modifications of the CF3-FTC chromophore architecture by increasing the donor strength will increase the transparency window in the visible. Further study involving modifications of acceptor and substitutions in the conjugated bridge are being carried out and will be reported in the future.

IV. Conclusions

We have assessed the efficiency of several DFT functionals for reproducing the experimental UV−vis charge-transfer and $\pi\rightarrow\pi^*$ absorption wavelengths of a set of push−pull chromophores containing the "tricyanofuran" (TCF) motif in the acceptor. These chromophores are thermally stable and exhibit excellent nonlinear optical properties.

Since our focus is to investigate optical transparency in the visible, the positions of major absorption peaks and their relative strengths are discussed in this work. We find that the PBE or SVWN give very poor estimates of absorption peaks, while hybrid functionals such as B3LYP and particularly PBE0 provide more consistent results for the molecules studied here. The performance of PBE0 is even more impressive considering its lack of empirical parametrization.

The long-range DFT Coulomb attenuation functionals offer a possibility to improve the accuracy of computed spectra by adjusting the attenuation parameter, γ. We have implemented the recently developed long-range-corrected Baer-Neuhauser-Livshits (BNL) exchange-correlation functional in the NWChem program and compared its performance with the popular CAM-B3LYP functional. The accuracy of UV−vis results is best at low values of attenuation parameters (γ) for both the BNL and CAM-B3LYP functionals. We observed that the optimal value of γ is different for the charge-transfer- and π- π^*-excitations. We recommend the BNL (with $\gamma = 0.1$) and PBE0 methods as they capture the charge-transfer states particularly well, although the $\pi-\pi^*$-excitations are computed less accurately depending on the specific system. This recommendation of the BNL method is valid for TCF chromophores and may not be applicable to other types of push−pull chromophores. In fact, transferability of both CAM-B3LYP and BNL methods is somewhat limited as their performance depends strongly on the value of attenuation parameter. "A priori" prediction of optimal value of γ for the BNL method remains an outstanding problem that only recently has been addressed, e.g. in calculations of intermolecular charge transfer complexes.[42] The value of γ optimized for the excitation energies is typically a bad choice when used for ground-state properties.[23,53,54] Recently Rohrdanz et al.[55] proposed a unified LC-DFT-type functional that shows promising accuracy in both ground and excited state calculations. Performance of this functional in calculations for the TCF chromophores is as yet unknown.

One of the goals of this work is to validate a practical, computational approach to predict spectra of chromophores in a realistic polymer environment. Chromophores may have side-chains linking a molecule to a polymer and exhibit numerous conformational states close in energy. Overall, a dielectric environment with a low dielectric constant may be expected. We confirm that the color of chromophores is not significantly affected by increasing aliphatic chain length linking a chromophore to a polymer. Unlike the strong dependence of the molecular hyperpolarizability on various conformers, we find that the optical absorption spectra are much less sensitive. The Boltzman statistical average of absorption peaks for various conformers is within ~10 nm of the spectra of the most stable conformer. This allows us

Optical Spectra of TCF-Chromophores

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2845**

to formulate an approximate, yet fairly accurate and practical approach. The lowest-energy-conformer is identified and its geometry optimized utilizing B3LYP in the gas phase, and the spectra are subsequently calculated using TD-DFT with an appropriate solvation model. The procedure outlined in this paper based on the BNL ($\gamma = 0.1$) functional in the COSMO solvent environment yields consistently good results with the largest error found to be ~40 nm in the case of $\pi-\pi^*$-excitations.

We have studied the effect of increased donor strength on absorption spectra for the CF3-FTC chromophore architecture with amino-phenyl donors. Substituting methoxy groups with methylamino groups, we succeeded in pushing the charge-transfer peak to the far-red. However, excitations of $\pi-\pi^*$-character are strengthened and even further red-shifted. Therefore, it is unlikely that the modifications of the CF3-FTC chromophore architecture by increasing the donor strength alone will increase the transparency window in the visible.

Supporting Information Available: Optimized structures. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Andrekson, P. A.; Westlund, M. *Laser Photonics Rev.* **2007**, *1*, 231–248.

(2) Pereverzev, Y. V.; Gunnerson, K. N.; Prezhdo, O. V.; Sullivan, P. A.; Liao, Y.; Olbricht, B. C.; Akelaitis, A. J.; Jen, A. K.-Y.; Dalton, L. R. *J. Phys. Chem. C* **2008**, *112*, 4355–4363, and references therein.

(3) Robinson, B. H.; Dalton, L. R.; Harper, A. W.; Ren, A.; Wang, F.; Zhang, Z.; Todorova, G.; Lee, M.; Aniszfeld, R.; Garner, S.; Chen, A.; Steier, W. H.; Houbrecht, S.; Persoons, A.; Ledoux, I.; Zyss, J.; Jen, A. K.-Y. *Chem. Phys.* **1999**, *245*, 35–50.

(4) (a) Zhang, C.; Dalton, L. R.; Oh, M.-C.; Zhang, H.; Steier, W. H. *Chem. Mater.* **2001**, *13* (9), 3043. (b) Jen, A.; Luo, J.; Kim, T.-D.; Chen, B.; Jang, S.-H.; Kang, J.-W.; Tucker, N. M.; Hau, S.; Tian, Y.; Ka, J.-W.; Haller, M.; Liao, Y.; Robinson, B.; Dalton, L.; Herman, W. *Proc. SPIE* **2005**, *5935*, 593506–1.

(5) Guenthner, A. J.; Wright, M. E.; Fallis, S.; Lindsay, G. A.; Petteys, B. J.; Yandek, G. R.; Zang, D. Y.; Sanghadasa, M.; Ashley, P. R. *Proc. SPIE* **2006**, *6331*, 63310M–1.

(6) Cheng, Y.-J.; Luo, J.; Hau, S.; Bale, D. H.; Kim, T.-D.; Shi, Z.; Lao, D. B.; Tucker, N. M.; Tian, Y.; Dalton, L. R.; Reid, P. J.; Jen, A. K.-Y. *Chem. Mater.* **2007**, *19*, 1154–1163.

(7) Davies, J. A.; Elangovan, A.; Sullivan, P. A.; Olbricht, B. C.; Bale, D. H.; Ewy, T. R.; Isborn, C. M.; Eichinger, B. E.; Robinson, B. H.; Reid, P. J.; Li, X.; Dalton, L. R. *J. Am. Chem. Soc.* **2008**, *130*, 10565.

(8) Kanis, D. R.; Ratner, M. A.; Marks, T. J. *Chem. Rev.* **1994**, *94*, 195–242.

(9) Gross, E. K. U.; Dobson, J. F.; Petersilka, M. *Density Functional Theory*; Springer: Berlin, 1996; Vol. 181.

(10) Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439.

(11) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.

(12) (a) Jacquemin, D.; Perpete, E.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105. (b) Jacquemin, D.; Perpete, E.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135.

(13) Magyar, R. J.; Tretiak, S. *J. Chem. Theory Comput.* **2007**, *3*, 976–987.

(14) Champagne, B.; Perpete, E. A.; Jacquemin, D.; Gisbergen, S. J. A.; Baerends, E.-J.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Phys. Chem. A* **2000**, *104*, 4755–4763.

(15) Dierksen; Grimme, S. *J. Phys. Chem. A* **2004**, *108*, 10225.

(16) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007–4016.

(17) Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697–12699.

(18) Savin, A. *Recent Advances in Density Functional Methods: Part I*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p 129.

(19) Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151.

(20) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *8*, 3540.

(21) Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.

(22) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(23) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.

(24) Chiba, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2006**, *124*, 144106.

(25) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, E.; Windus, T. L.; Hammond, J.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Wu, Q.; Van Voorhis, T.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 5.1*; Pacific Northwest National Laboratory: Richland, Washington 99352-0999, USA, 2007.

(26) (a) Hayden, L. M.; Sauter, G. F.; Ore, R. F.; Pasillas, P. L.; Hoover, J. M.; Lindsay, G. A.; Henry, R. A. *J. Appl. Phys.* **1990**, *88*, 456–465. (b) UV-vis spectrum of TCV in 1,2-dichloroethane was measured at the Army Research Laboratory.

(27) Isborn, C. M.; Leclercq, A.; Vila, F. D.; Dalton, L. R.; Bredas, J. L.; Eichinger, B. E.; Robinson, B. H. *J. Phys. Chem. A* **2007**, *111*, 1319–1327.

(28) Kinnibrugh, T.; Bhattacharjee, S.; Sullivan, P.; Isborn, C.; Robinson, B. H.; Eichinger, B. E. *J. Phys. Chem. B* **2006**, *110*, 13512–13522.

(29) Leclercq, A.; Zojer, E.; Jang, S.-H.; Barlow, S.; Geskin, V.; Jen, A. K.-Y.; Marder, S. R.; Bredas, J. L. *J. Chem. Phys.* **2006**, *124*, 044510−044510.

(30) Schmidt, K.; Barlow, S.; Leclercq, A.; Zojer, E.; Jang, S.-H.; Marder, S. R.; Jen, A. K.-Y.; Bredas, A.-L. *J. Mat. Chem.* **2007**, *17*, 2944–2949.

(31) 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 E.01*; Gaussian, Inc.: Wallingford, CT, 2004.

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

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

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

(35) (a) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094. (b) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43.

(36) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.

(37) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799.

(38) Parr, R. G.; Yang, W. *Density-functional theory of atoms and molecules*; Oxford Univ. Press: Oxford, 1989.

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

(40) (a) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036. (b) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(41) Becke, A. *J. Chem. Phys.* **1993**, *8*, 1372–1377.

(42) (a) Livshits, E.; Baer, R. *J. Phys. Chem. A* **2008**, *112*, 12789–12791. (b) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818.

(43) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *314*, 291–299.

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

(45) Von Arnim, M.; Ahlrichs, R. *J. Comput. Chem.* **1998**, *19*, 1746–1757.

(46) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560–571.

(47) Andrade do Monte1, S.; Muller, T.; Dallos, M.; Lischka, H.; Diedenhofen, M.; Klamt, A. *Theor. Chem. Acc.* **2004**, *111*, 78–89.

(48) Klamt, A.; Jonas, V.; Berger, T.; Lohrenz, J. C. W. *Phys. Chem. A* **1998**, *102*, 5074–5085.

(49) Stefanovich, E. V.; Truong, T. N. *Chem. Phys. Lett.* **1995**, *244*, 65–74.

(50) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(51) Pearl, G. M.; Zerner, M. C.; Broo, A.; McKelvey, J. *J. Comput. Chem.* **1998**, *19*, 781–796.

(52) Wong, B. M.; Cordaro, J. G. *J. Chem. Phys.* **2008**, *129*, 214703.

(53) Rohrdanz, M.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107.

(54) Peach, M. J. G.; Cohen, A. J.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4543–4549.

(55) Rohrdanz, M.; Martins, K. M.; Herbert, J. M. *J. Chem. Phys.* **2008**, *130*, 054112.

(56) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.

(57) Jacquemin, D.; Perpete, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 094102.

CT900231R

JCTC Journal of Chemical Theory and Computation

Simulated Solute Tempering

Robert Denschlag, Martin Lingenheil, Paul Tavan, and Gerald Mathias*

Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität, Oettingenstrasse 67, 80538 München, Germany

Abstract: For the enhanced conformational sampling in molecular dynamics (MD) simulations, we present "simulated solute tempering" (SST) which is an easy to implement variant of simulated tempering. SST extends conventional simulated tempering (CST) by key concepts of "replica exchange with solute tempering" (REST, Liu et al. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749). We have applied SST, CST, and REST to molecular dynamics (MD) simulations of an alanine octapeptide in explicit water. The weight parameters required for CST and SST are determined by two different formulas whose performance is compared. For SST only one of them yields a uniform sampling of the temperature space. Compared to CST and REST, SST provides the highest exchange probabilities between neighboring rungs in the temperature ladder. Concomitantly, SST leads to the fastest diffusion of the simulation system through the temperature space, in particular, if the "even-odd" exchange scheme is employed in SST. As a result, SST exhibits the highest sampling speed of the investigated tempering methods.

Introduction

The generation of equilibrium ensembles for macromolecules by all-atom Monte Carlo (MC) or molecular dynamics (MD) simulations is a challenging task due to the huge computational effort which is generally necessary to guarantee ergodic sampling of relevant observables. The required simulation time depends on the number and on the depths of the local minima in the free energy landscape because here the simulation may get trapped for extended periods of time. If the barriers between the minima are mainly of enthalpic nature, generalized ensemble tempering techniques enable faster barrier crossings and, therefore, alleviate the sampling problem.[1,2]

Two important generalized ensemble tempering algorithms are simulated tempering (ST) and replica exchange (RE).[2] The RE method in its original form[3−5] employs several copies (replicas) of the investigated system at different temperatures $T_0 < T_1 < \cdots < T_{N-1}$. Within this temperature ladder, temperature exchanges between the replicas are periodically attempted. The corresponding probability for an exchange is given by a Metropolis criterion.[6] On the other hand, in simulated tempering,[7,8] only one "replica" diffuses along the temperature ladder, where temperature changes are determined by a Metropolis criterion slightly modified with respect to RE. Although both methods are closely related,[9] RE has attracted much more attention than ST, which is indicated by the large number of RE variants that have been suggested.[10−17] Furthermore, the RE efficiency has been the subject of many studies,[18−23] and numerous RE applications to macromolecules have been presented.[24−33] The main reason for the apparent neglect of ST is that this approach requires estimates for certain *a priori* unknown parameters, the so-called weights, to ensure a uniform sampling of the ensembles at all temperatures. In contrast, RE automatically guarantees a uniform sampling of all temperatures and is, therefore, much simpler to control.

Without prior knowledge on the properties of the simulated system an unbiased tempering algorithm should uniformly cover the chosen temperature space to generate an enhanced statistics at the temperature of interest, which usually is the lowest temperature T_0. Therefore, the average time to shuttle a replica between T_0 and the maximal temperature T_{N-1} should be by orders of magnitude shorter than the simulation time. Addressing this issue, Abraham and Gready recently exam-

* Corresponding author phone: +49-89-2180-9228; fax: +49-89-2180-9202; e-mail: gerald.mathias@physik.uni-muenchen.de. Corresponding author address: LMU München, Lehrstuhl für BioMolekulare Optik, Oettingenstrasse 67, D-80538 München, Germany.

ined more than forty published RE simulations for their ability to take full advantage of the tempering.[34] In most cases, they found that the simulation times were too short for a sufficiently frequent shuttling of the replicas along the temperature ladder. The reason for large shuttling times is known: This time usually scales with the square of the number of temperatures,[35] which in turn grows with the square root of the number of degrees of freedom[3] (DOF) in the simulation system. Thus, very long round trip times are expected for simulation systems with many DOF. Typical examples for such systems are macromolecules in explicit solvent. Note that such systems may additionally undergo slow phase transitions, like folding and unfolding, which may drastically increase round trip times.

Various strategies aiming at increasing the shuttling frequency were suggested for RE,[14−16,35−38] which should all be transferable to ST. A first class of strategies targets the optimization of the parameters that characterize the setup of an RE simulation. For example, Sindhikara et al.[38] recently recommended that the times between exchange trials should be as small as possible, which, however, has been disputed.[34] Additionally, rules for optimizing the temperature ladders were suggested.[35,39] A second class of strategies tries to reduce the number of DOF that are relevant for the exchange criterion, for example some strategies switch from an explicit to an implicit solvent description for the exchange trials[16] or by tempering only the areas of interest.[15] With such a reduced number of DOF, the temperature steps within the ladder can be chosen larger and, thus, a given temperature range can be covered by less rungs. As a result, the round trip times are drastically shortened. However, the quoted methods of the second class are not rigorous in terms of statistical physics because here the ensemble is modified in a somewhat arbitrary fashion. Recently, Liu et al.[14] have presented a rigorous strategy called "replica exchange with solute tempering" (REST), which largely eliminates the influence of the solvent DOF on the exchange probabilities by a temperature dependent scaling of the Hamiltonian. In this approach only the Hamiltonian at the target temperature remains unscaled and renders the desired physically meaningful ensemble. This restriction is, however, of minor importance for many applications.

An additional strategy to achieve increased shuttling frequencies is the optimal choice of the tempering method itself. For example, with accurately known weight parameters, ST is more efficient than RE because it provides larger acceptance probabilities on the same temperature ladder.[40−42] At given conditions one may equivalently state that ST requires less rungs in the temperature ladder than RE if both techniques are tuned to the same average acceptance probabilities.

The weights necessary for ST can be estimated by short preparatory simulations bearing the risk, however, that these weights are of insufficient accuracy. Addressing this issue, a recent study by Park and Pande[9] suggests that controlling an ST simulation may be less difficult than previously assumed by Mitsutake and Okamoto.[43] In contrast to the latter authors Park and Pande[9] did not stick to the rough estimates for the weights derived from a set of short preparatory simulations but instead updated the weights during the subsequent ST production run.

Here, inspired by the works of Park and Pande[9] and of Liu et al.,[14] we suggest a variant of ST called simulated solute tempering (SST), which shares key concepts with REST and its sequential variant SREST.[13] As we will demonstrate, SST has the advantages of a most simple implementation and of reducing the required number of rungs within the temperature ladder. Note, that SST should be easily transferrable to hybrid methods which combine different tempering techniques[43] or tempering with other enhanced sampling methods.[44]

We start in "Theory and Methods" with an introduction to replica exchange and its variant REST, which employs a temperature dependent scaling of the Hamiltonian. Then, we present SST along with two procedures for calculating the weights and corresponding update schemes. Subsequently, these techniques are applied to an alanine octapeptide (8ALA) solvated in water. The enthalpic barriers of 8ALA are small, and, hence, the benefit of tempering techniques is limited.[22] However, the small barriers provide fast conformational sampling which makes 8ALA ideally suited to compare different tempering strategies among each other with high statistical accuracy. In particular, we investigate the sampling efficiency and convergence of conventional ST (CST), SST, and REST. In addition, we test two temperature exchange schemes for SST to improve the method further. After the presentation and discussion of the results, we conclude the paper summarizing the key messages.

Theory and Methods

We begin by sketching the replica exchange method and the concept of solute tempering, which will lead us, when combined with simulated tempering, to the SST method.

Replica Exchange. Within both conventional temperature RE[5] (CRE) and REST,[14] N copies (replicas) of the system are simulated at temperatures $T_0 < T_1 < \cdots < T_{N-1}$ and sample the associated canonical ensembles. The set of replicas constitutes a so-called generalized ensemble. After predefined time intervals a temperature exchange between pairs of replicas is tried. A Metropolis criterion[6] determines the exchange probability

$$P_{ij} = \min[1, \exp(\Delta_{ij})] \qquad (1)$$

between replicas at T_i and T_j with

$$\Delta_{ij} = \beta_i[E_i(\mathbf{x}_j) - E_i(\mathbf{x}_i)] + \beta_j[E_j(\mathbf{x}_i) - E_j(\mathbf{x}_j)] \qquad (2)$$

to preserve the canonical ensembles. Here, $E_k(\mathbf{x}_l)$ is the value of the potential energy function associated with the temperature T_k, which is evaluated at a configuration $\mathbf{x}_l$ resulting from the sampling at T_l; $\beta_k = 1/k_B T_k$ is the inverse temperature where k_B is the Boltzmann constant. For CRE the potential energy is independent of the temperature, and Δ_{ij} reduces to

$$\Delta_{ij} = (\beta_i - \beta_j) \cdot [E(\mathbf{x}_i) - E(\mathbf{x}_j)] \qquad (3)$$

In the case of REST, in contrast, the potential energy becomes temperature dependent and has the form

Simulated Solute Tempering

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2849**

$$E_k(\mathbf{x}) = \lambda_{k,0}E^{\mathrm{pp}}(\mathbf{x}) + \lambda_{k,1}E^{\mathrm{ps}}(\mathbf{x}) + \lambda_{k,2}E^{\mathrm{ss}}(\mathbf{x}) \quad (4)$$

where E^{pp}, E^{ps}, and E^{ss} are the solute−solute, solute−solvent, and solvent−solvent parts of the potential energy function at the target temperature T_0 of the sampling; $\lambda_{k,h}$ are parameters depending on the temperature T_k. We choose

$$\lambda_{k,0} = 1, \ \lambda_{k,1} = \sqrt{\beta_0/\beta_k}, \ \lambda_{k,2} = \beta_0/\beta_k \quad (5)$$

where $\lambda_{k,1}$ is the geometric mean of $\lambda_{k,0}$ and $\lambda_{k,2}$ instead of the arithmetic mean $(\beta_0 + \beta_k)/2\beta_k$ originally proposed by Liu et al.[14] With this choice, the required scaling of the electrostatic energy and of the corresponding forces at T_k can be achieved by simply scaling the partial charges of the solvent by a factor $(\beta_0/\beta_k)^{1/2}$. Similar considerations hold for the Lennard-Jones interactions. Thus, this choice of the $\lambda_{k,i}$ is conveniently implemented.

The advantage of the REST approach becomes apparent after a few algebraic operations. Inserting eqs 4 and 5 into eq 2 one finds

$$\Delta_{ij} = (\beta_i - \beta_j)[E^{\mathrm{pp}}(\mathbf{x}_i) - E^{\mathrm{pp}}(\mathbf{x}_j)] + (\sqrt{\beta_0\beta_i} - \sqrt{\beta_0\beta_j})[E^{\mathrm{sp}}(\mathbf{x}_i) - E^{\mathrm{sp}}(\mathbf{x}_j)] \quad (6)$$

Thus, the difference Δ_{ij}, which determines the acceptance probability eq 1, is exclusively calculated from the solute−solute and solute−solvent energies, whereas the potential energy E^{ss} of the solvent cancels.

One can quantify the benefit of REST compared with CRE by estimating the number of rungs eliminated from the temperature ladder. Assuming that the solvent DOF do not contribute to the exchange probability at all, the ratio $N_{\mathrm{REST}}/N_{\mathrm{CRE}}$ of the required rungs is estimated by the lower limit $(n_p/(n_s + n_p))^{1/2}$, where n_s and n_p count the DOF of the solvent and solute, respectively. In the following we discuss how simulated tempering can further reduce the required number of rungs.

Simulated Tempering. In simulated tempering[7,8] (ST), a single system is simulated at a temperature T_i, which belongs to a given temperature ladder $T_0 < \cdots < T_{N-1}$. After given time intervals it is checked whether the system temperature T_i can be changed to T_j, where j is usually $i \pm 1$. For $i \in \{0, N-1\}$ the transition to $j = -1$ or $j = N$ is rejected. For other transitions $i \rightarrow j$, the acceptance probability[41]

$$P_{ij} = \min[1, \exp(\Delta_{ij})] \quad (7)$$

with

$$\Delta_{ij} = [\beta_i E_i(\mathbf{x}) - w_i] - [\beta_j E_j(\mathbf{x}) - w_j] \quad (8)$$

represents a Metropolis criterion similar to that of RE. The weights w_k introduced in eq 8 are commonly set to the configurational parts $\beta_k\tilde{F}_k = -\ln \int \exp[-\beta_k E(\mathbf{x})]d\mathbf{x}$ of the dimensionless free energies $\beta_k F_k$ of the simulation system at the temperatures T_k.[7,8] This choice leads to a uniform sampling of all rungs within a given temperature ladder because, in the ergodic limit, the expected ratio $\rho_k \equiv t_k/t$ of the time t_k spent by the simulation at temperature T_k to the total sampling time t is given by the Boltzmann factor of the generalized ensemble

$$\lim_{t\to\infty} \rho_k = \frac{\exp[-(\beta_k\tilde{F}_k - w_k)]}{\sum_j \exp[-(\beta_j\tilde{F}_j - w_j)]} \quad (9)$$

leading to $\lim_{t\to\infty} \rho_k = 1/N$ for $w_j = \beta_j\tilde{F}_j$. Since the w_k are *a priori* unknown, one usually tries to estimate these weights from short preparatory simulations. Note that the w_j can be chosen differently, if a nonuniform sampling is desired.[42]

In conventional simulated tempering (CST), the potential energy function $E_k(\mathbf{x})$ is independent of the temperature T_k, and eq 8 reduces to

$$\Delta_{ij} = (\beta_i - \beta_j)E(\mathbf{x}) - (w_i - w_j) \quad (10)$$

For the new SST method introduced here, we transfer the solute tempering concept of REST to ST and use the energy function given by eqs 4 and 5. Inserting these equations into eq 8 yields

$$\Delta_{ij} = (\beta_i - \beta_j)E^{\mathrm{pp}}(\mathbf{x}) + (\sqrt{\beta_0\beta_j} - \sqrt{\beta_0\beta_i})E^{\mathrm{ps}}(\mathbf{x}) - (w_i - w_j) \quad (11)$$

Thus, for SST the difference Δ_{ij} is only calculated from the solute−solute and solute−solvent energies, while the potential energy E^{ss} of the solvent cancels. As we will show below the solvent−solvent contributions cancel as well in the computation of the weight differences $w_i - w_j$.

At first glance, ST (CST/SST) seems less attractive than RE (CRE/REST) because ST requires the *a priori* unknown weights w_i. However, ST provides larger average acceptance probabilities than RE[40−42] for a given temperature ladder because RE requires a simultaneous exchange of two replicas, whereas only one replica has to be considered for ST. As a result, ST needs only $1/\sqrt{2}$ times the number of rungs in the temperature ladder than RE to cover a given temperature range with the same acceptance probabilities.[39] Now, we turn to different approaches to determine the required weights w_i.

Determination of the Weights w_k. As we have seen above, a uniform sampling of the various temperatures T_k requires that the weights w_k in eq 8 are the configurational parts $\beta_k\tilde{F}_k$ of the dimensionless free energies, which can be estimated from preparatory simulations. For CST, Park and Pande have presented a formula which yields surprisingly good estimates of the w_k at a negligible computational effort (e.g., by executing a single 10 ps MD simulation at each T_k).[9] These authors replaced the potential energy $E(\mathbf{x})$ in eq 10 by the average potential energy $\langle E\rangle_i$ at T_i to get a "typical" $\Delta_{ij}^{\mathrm{typ}}$ and demanded that $\Delta_{ij}^{\mathrm{typ}} = \Delta_{ji}^{\mathrm{typ}}$, which leads to the estimate

$$w_j - w_i \approx (\beta_j - \beta_i)\frac{\langle E\rangle_i + \langle E\rangle_j}{2} \quad (12)$$

and which has been further substantiated by Park.[41] In the Appendix we show an alternative derivation of the "trapezoid rule" eq 12 that is based on the assumption of a constant heat capacity C_V and that $\ln[1 + (T_j - T_i)/T_i]$ can be approximated by $(T_j - T_i)/T_i$.

Replacing the temperature dependent potential energies $E_i(\mathbf{x})$ and $E_j(\mathbf{x})$ in eq 8 by averages $\langle E_i \rangle_i$ and $\langle E_j \rangle_i$ transfers the trapezoid rule of CST to SST. Here, $\langle E_j \rangle_i$ is the energy function E_j evaluated for and averaged over the configurations sampled at T_i. Analogous to CST, also these SST averages yield "typical" differences Δ_{ij}^{typ}. Equation 5 and $\Delta_{ij}^{\text{typ}} = \Delta_{ji}^{\text{typ}}$ lead to

$$
\begin{aligned}
w_j - w_i \approx \ & \frac{(\beta_j - \beta_i)(\langle E^{pp} \rangle_i + \langle E^{pp} \rangle_j)}{2} \\
& + \frac{(\sqrt{\beta_0 \beta_j} - \sqrt{\beta_0 \beta_i})(\langle E^{ps} \rangle_i + \langle E^{ps} \rangle_j)}{2}
\end{aligned}
\tag{13}
$$

As mentioned in the section "Simulated Tempering", $w_j - w_i$ does not depend on the solvent−solvent interactions E^{ss}. Note that one can choose $w_0 = 0$ because only differences $w_j - w_i$ matter in eq 8.

Whereas the computations of the weights w_i in CST and SST by the trapezoid rules eqs 12 and 13, repectively, are approximate, the relation

$$
\exp(-w_i) = \sum_{k=0}^{N-1} \sum_{t=1}^{n_k} \frac{\exp\{-\beta_i \sum_{j=0}^{L-1} \lambda_{i,j} E^j[\mathbf{x}_k(t)]\}}{\sum_{m=0}^{N-1} n_m \exp\{w_m - \beta_m \sum_{j=0}^{L-1} \lambda_{m,j} E^j[\mathbf{x}_k(t)]\}}
\tag{14}
$$

presented earlier by Kumar et al.[45] for the "weighted histogram analysis method" (WHAM) is "exact" for the already sampled ensemble and provides an unbiased estimator for the true dimensionless free energies.[46] Here, N is again the number of temperatures T_k, n_k is the number of configurations $x_k(1),...,x_k(n_k)$ sampled at T_k, and L counts potential energy terms contributing to the Hamiltonian. For CST there is only one such term and one has $L = 1$, $E^0(\mathbf{x}_k) \equiv E(\mathbf{x}_k)$, and $\lambda_{i,0} = 1$. The Hamiltonian of SST eq 4 distinguishes $L = 3$ energy contributions $E^0(\mathbf{x}_k) \equiv E^{pp}(\mathbf{x}_k)$, $E^1(\mathbf{x}_k) \equiv E^{ps}(\mathbf{x}_k)$, and $E^2(\mathbf{x}_k) \equiv E^{ss}(\mathbf{x}_k)$. With the $\lambda_{i,j}$ chosen as given by eq 5 one finds that E^{ss} cancels in eq 14 like it did in the trapezoid rule eq 13. Equation 14 has to be solved self-consistently and yields successively more and more accurate weights as the statistics is improved by an ongoing sampling. Note that Mitsutake and Okamoto[40,43] previously suggested to compute the free energies required for ST through WHAM equations, which are based on energy histograms. In contrast, eq 14 computes the free energies directly from the sampled energies and, therefore, avoids the errors introduced by the histogram discretization.[45−47]

The WHAM formula eq 14 can be used to identify the errors Δw_i of the trapezoid rules eq 12 for CST and eq 13 for SST. In the limit of ergodic sampling, the errors $\Delta w_i = w_i - w_i^{\text{exact}}$ yield through eq 9 the ratios $\lim_{t \to \infty} \rho_i = \exp(\Delta w_i)/\sum_j \exp(\Delta w_j)$. These ratios will deviate from $1/N$ and, therefore, measure deviations from the desired uniformity of the sampling along the ladder of temperatures T_i. To measure this deviation in our simulations, we introduce the quantities

$$
\chi_i \equiv N \frac{t_i}{t}
\tag{15}
$$

A uniform sampling corresponds to $\chi_i = 1.0$ at all T_i. Thus, the χ_i enable easy comparisons of the sampling uniformity achieved with differently sized temperature ladders. If the errors Δw_i are known, the long time limit of χ_i is

$$
\hat{\chi}_i = N \frac{\exp(\Delta w_i)}{\sum_{j=0}^{N-1} \exp(\Delta w_j)}
\tag{16}
$$

For finite simulations, the χ_i exhibit statistical fluctuations which depend on the number of rungs N in the temperature ladder, on the average exchange probabilities $\bar{P}_{ij}$, and on the total number of exchange trials. For each set of these parameters, one can model an actual ST simulation by a computationally inexpensive MC simulation of a random walk along the N rungs of the temperature ladder. We will use large numbers of such MC simulations to estimate the standard deviations σ_i of the χ_i for the respective simulations. These values show to what extent one may expect convergence of sampling uniformity.

Update Schemes for the Weights w_i. Initial guesses for the weights required in ST can be obtained from short preparatory simulations using the formulas presented in the previous paragraph. The correspondingly limited statistical accuracy of the initial weights may entail a strongly nonuniform sampling along the temperature ladder in subsequent ST production runs. However, one may improve the initial guesses by utilizing the information accumulated in the course of the production run. For this purpose different adaptation schemes were suggested.[48−51]

We used a procedure based on the following considerations: Up to the first update the sampling along the temperature ladder is expected to be far off from uniformity. Consequently, a poor statistics of the potential energy distribution is obtained at some temperature rungs. To avoid an impact of this bad statistics on the estimated weights, we determine the first update by $w_i^{\text{new}} = w_i^{\text{old}} + \ln(t_0/t_i)$ where we set the t_i for rungs that have not been visited at all to a full exchange period. Thus, badly sampled rungs will be preferentially sampled until the next update step. By construction this strategy leads to a uniform sampling but may suffer from slow convergence. Therefore, we subsequently switch to a periodical recomputation of the weights either by the trapezoid rule (eqs 12 and 13) or by the WHAM formula (eq 14). In these recomputations, which are executed after each nanosecond of simulation, we exclusively consider the data from the production run and discarded those of the preparatory simulation.

Exchange Scheme. Having established the determination of the weights steering the exchange probabilities, we now sketch the exchange algorithms employed for ST. A straightforward exchange procedure is to choose randomly between an upward or downward exchange trial with probabilities of 50%. We call this exchange procedure "stochastic even/odd" (SEO) scheme because the corresponding exchange scheme for replica exchange is a stochastic instead of a deterministic choice between two groups of replica pairs.[52] The first group contains all "even" pairs (T_{2n}, T_{2n+1}) and the second one all "odd" pairs (T_{2n-1}, T_{2n}).

Simulated Solute Tempering

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2851**

Table 1. Overview of Simulations Conducted[a]

label	trajectory time span/ns	temperature range/K	no. of rungs	solvent scaling	determination of weights	exchange scheme
CRE	18 × 0.1	300−500	18	no	-	DEO
CST	27	300−500	18	no	trapezoid	SEO
REST/A	4 × 0.1	300−500	4	yes	-	DEO
REST/B	5 × 0.1	300−500	5	yes	-	DEO
REST/C	4 × 12	300−500	4	yes	-	DEO
SST/A	27	300−500	4	yes	trapezoid	SEO
SST/B	27	300−500	4	yes	WHAM	SEO
SST/C	12	300−500	4	yes	WHAM	DEO
SST/D	12	300−500	5	yes	WHAM	DEO

[a] The weights for the ST simulations have been determined either by the trapezoid rule, eq 12 for CST and eq 13 for SST, respectively, or by the WHAM formula eq 14.

However, the standard exchange scheme used in replica exchange simulations is characterized by alternate exchange trials between these two groups of replica pairs,[3,34,52,53] which we have called the "deterministic even/odd" (DEO) scheme.[52] Formally, one can express the DEO exchange scheme by a relation which combines the involved temperature indices i and i' at the exchange attempt step s by $i' = i + (-1)^{i+s}$. In the framework of simulated tempering DEO alternately tries to shift the single replica at T_i to T_{i+1} (upward) or T_{i-1} (downward). In the case of a successful exchange, however, the previous exchange direction (upward or downward) is maintained for the next exchange trial and so forth until a temperature exchange fails.

We apply both exchange schemes to investigate their influence on the diffusion of the system through the temperature space. This diffusion can be measured in terms of the average round trip time τ required to travel from T_0 to T_{N-1} and back to T_0. For the SEO scheme applied to a temperature ladder with uniform average acceptance probabilities $\bar{P}_{ij} = \bar{P}$, τ in units of the time between exchange trials is related to the average acceptance probability $\bar{P}$ by[35,52,54,55]

$$\tau_{SEO} = 2N(N - 1)/\bar{P} \tag{17}$$

For DEO, τ is given by

$$\tau_{DEO} = \tau_{SEO}(1 - \bar{P}) \tag{18}$$

in the limit of large N.[52] Assuming this limit, DEO always provides shorter round trip times than SEO or, equivalently, higher round trip rates τ^{-1}. Thus, the question is to what extent this expectation is confirmed for finite ladder sizes N.

Simulation System and Force Field. We have used MD simulations of a poly alanine octapeptide (8ALA), saturated with an acetyl group at the N-terminus and an *N*-methyl group at the C-terminus, to investigate the benefits of the various algorithms introduced above. The peptide was described by the CHARMM22 force field[56] and solvated in a periodic orthorhombic dodecahedron of 18 Å inscription radius containing 1112 water molecules. For the water molecules we employed the transferable three-point intermolecular potential (TIP3P),[57] modified as suggested by MacKerell et al.[56] for usage with CHARMM22. The initial 8ALA structure was generated using the Molden software[58]

by setting the backbone dihedral angles to the values $\phi = -58°$ and $\psi = -47°$ to form an ideal α-helix.

MD Simulation Techniques. The software package EGO-MMVI[59] was used for all MD simulations. The electrostatic interactions were treated combining structure-adapted multipole expansions[60] with a moving-boundary reaction-field approach.[59] Here, the cutoff radius for the explicit evaluation of the electrostatic interactions was 18 Å. Beyond this radius, a dielectric continuum was assumed with a static dielectric constant $\varepsilon_s = 80$. The explicit van der Waals interactions were calculated up to a distance of 10 Å and at larger distances a mean-field approach was applied.[61] A multiple-time-step integration scheme[62] with a fastest time step of 2 fs was used. For bonds that include hydrogen atoms, the corresponding bond lengths were constrained using the M-SHAKE algorithm[63] with relative tolerance of 10^{-6}.

System Preparation. Our simulation system was equilibrated for 100 ps with two Berendsen thermostats[64] (coupling times 0.1 ps) separately keeping the solute and the solvent at 300 K. Additionally, a Berendsen barostat[64] (coupling time 1 ps) steered the system to ambient pressure (1 bar). For the subsequent tempering runs we switched from an *NPT* to an *NVT* ensemble.

Simulation Runs. As listed in Table 1 we carried out three short replica exchange simulations [CRE and REST/(A,B)] serving to estimate the initial weights for the extended simulated tempering simulations CST and SST/A-D. For CRE and CST we used the temperature ladder 300 K, 308 K, 317 K, 326 K, 336 K, 346 K, 356 K, 367 K, 378 K, 390 K, 402 K, 415 K, 428 K, 442 K, 456 K, 470 K, 485 K, and 500 K. We found average acceptance probabilities $\bar{P}_{i,i+1}$ between 5% and 14% for CRE and between 21% and 31% for CST. For REST/(A,C) and SST/A-C we used the ladder 300 K, 350 K, 415 K, and 500 K. The corresponding $\bar{P}_{i,i+1}$ range from 5% to 12% (REST) and from 19% to 28% (SST). For REST/B and SST/D, the five temperatures 300 K, 340 K, 387 K, 440 K, and 500 K were used yielding $\bar{P}_{i,i+1}$ between 15% and 19% (REST) and between 38% and 41% (SST). Exchanges were tried every 0.5 ps in all simulations. Tables S1 and S2 in the Supporting Information provide details about the average exchange probabilities along the various ladders. These values were used for setting up the MC simulations mentioned further above.

The extended simulations REST/C, CST, and SST/(A,B) serve us for comparisons of methods and address, in particular, the applicability of different adaptation schemes

2852 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Denschlag et al.

to SST. Furthermore, using the simulations SST/C and SST/D we will study the effects of the chosen exchange scheme (SEO vs DEO) and of the (overall) average exchange probability $\bar{P} = \langle \bar{P}_{ij} \rangle$ on the round trip rates τ^{-1} and the sampling speeds.

Sampling Speed. The main objective of tempering methods is to enhance the sampling speed of the simulation. A corresponding measure for the sampling speed is given by an algorithm recently suggested by Lyman and Zuckerman,[65] which we will denote as LZA. LZA integrates the "volume" in configurational space sampled by a trajectory. The average volume sampled during a given simulation time span provides a measure for the sampling speed.

For 8ALA we define the conformational space by the eight dihedral angles ψ_i spanned by the backbone units $N^i - C^i - C_\alpha^i - N^i$. From a trajectory of the eight-dimensional tuples $(\psi_1,...,\psi_8)$, LZA randomly chooses one tuple and removes it from the trajectory together with all other tuples lying within the sphere of predefined radius r around the chosen tuple. This procedure is repeated until all tuples of the initial trajectory have been removed. The number of steps required is a dimensionless measure for the configurational volume V_c sampled by the trajectory. Because this algorithm is nondeterministic, it is repeated m times, and the corresponding average number n_{lza} of required steps is calculated. For our analyses we choose $r = 25°\sqrt{8}$ and $m = 50$. For a fair comparison between ST and RE, we compute the sampling speed per replica, i.e. one for ST and N for RE.

Results and Discussion

At the start of an ST simulation, initial estimates for the weight parameters w_i are needed. We determined these estimates from preparatory simulations using both the approximate trapezoid rule eq 12 and the asymptotically unbiased WHAM formula eq 14. Now, a first issue is the reliability of the trapezoid rule, which we check using the 100 ps CST simulation.

Reliability of the Trapezoid Rule for CST. Table 2 compares the initial CST weights w_i determined by the trapezoid rule eq 12 from the preparatory CRE simulation with the asymptotically unbiased values calculated by the WHAM formula eq 14. For all T_i the w_i obtained by the two formulas agree quite well. The errors Δw_i of eq 12 never exceed 12% of $k_B T_i$, and, correspondingly, the uniformity measures $\hat{\chi}_i$ are all close to 1.0. Hence, one expects a nearly uniform sampling even if the weights are determined by eq 12. Thus, adaptation schemes, which are based on the trapezoid rule and on the WHAM formula, respectively, should be nearly equivalent for the given system. In the CST simulation we, therefore, applied the trapezoid rule for periodical recomputation of the w_i.

Representative for the eighteen weights, Figure 1(a) shows the deviation of the weights w_8 and w_{17} from their initial values as a function of the simulation time. The exceptional first update $w_k^{new} = w_k^{old} + \ln(t_0/t_k)$, which can be seen as a special case of the update scheme proposed by Zhang and Ma,[51] sizably reduces both weights and reflects the nonuniform sampling within the preceding first nanosecond of the

Table 2. Weights Determined from the CRE Simulation[a]

i	T_i	w_i (trapezoid)	w_i (WHAM)	Δw_i	$\hat{\chi}_i$
0	300 K	0.0	0.0	−0.00	1.06
1	308 K	482.58	482.59	−0.01	1.05
2	317 K	991.97	991.99	−0.02	1.04
3	326 K	1468.66	1468.69	−0.03	1.03
4	336 K	1963.47	1963.49	−0.02	1.04
5	346 K	2425.14	2425.17	−0.03	1.03
6	356 K	2856.65	2856.68	−0.03	1.03
7	367 K	3299.53	3299.56	−0.03	1.03
8	378 K	3712.36	3712.41	−0.05	1.00
9	390 K	4131.64	4131.73	−0.09	0.97
10	402 K	4521.49	4521.56	−0.07	0.99
11	415 K	4913.99	4914.07	−0.08	0.98
12	428 K	5278.33	5278.44	−0.11	0.95
13	442 K	5642.42	5642.50	−0.08	0.98
14	456 K	5980.14	5980.21	−0.07	0.99
15	470 K	6294.02	6294.09	−0.07	0.99
16	485 K	6606.46	6606.55	−0.09	0.97
17	500 K	6896.68	6896.80	−0.12	0.94

[a] The weights w_i determined by the trapezoid rule eq 12 and by the WHAM formula eq 14 together with the deviations Δw_i and the correspondingly predicted (cf. eq 16) uniformity measures $\hat{\chi}_i$. The WHAM weights were employed as starting values for the CST simulation.
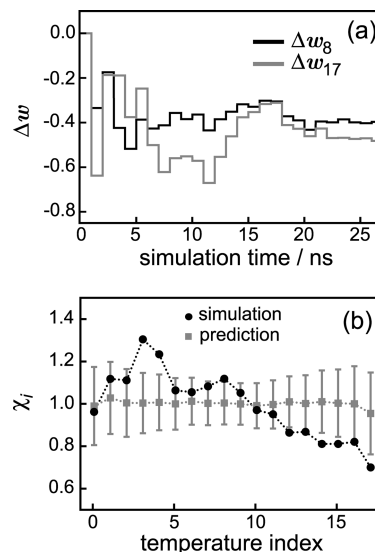

Figure 1. Uniformity of the temperature sampling in the CST simulation. (a) Time evolution of the weights w_8 and w_{17} with respect to their initial values. (b) Uniformity measures observed (χ_i, eq 15, circles) and predicted ($\hat{\chi}_i$, eq 16, squares) after 27 ns at the temperatures T_i. The standard deviations were estimated from MC trial simulations (see text for further details). The dotted lines serve as a guide for the eye.

CST simulation. Here, the temperatures $T_8 = 378$ K and $T_{17} = 500$ K apparently have been visited more frequently than $T_0 = 300$ K. The following updates, which rely on eq 12, lead to considerable changes of the weights, which, however, become smaller toward the end of the simulation. After 27 ns the weights seem to be converged within roughly ± 0.1. This is approximately the same magnitude of error as the one introduced by the trapezoid rule.

Figure 1(b) shows the measured (circles) and expected (squares) uniformity measures χ_i and $\hat{\chi}_i$ extracted from the last 25 ns of the CST simulation as functions of the temperature. In contrast, the uniformity data shown in Table

Simulated Solute Tempering

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2853**

2 had been extracted from the much shorter preparatory CRE simulation. According to eq 16 the observables $\hat{\chi}_i$ reflect the average deviations Δw_i between the trapezoid weights used during the CST simulation and WHAM weights calculated *a posteriori* from that simulation. As one sees in the figure, the expectation values $\hat{\chi}_i$ are close to one demonstrating that the trapezoid rule induces only small errors into the weights. These data confirm the claim[9] that the trapezoid rule is appropriate for choosing the weights in conventional ST simulations.

Despite the expected nearly uniform sampling of the sampling along the temperature ladder the values χ_i measured for the CST simulation deviate substantially from one, yielding a root-mean-square deviation (RMSD) from uniformity of 16%. Because the weights are calculated with a reasonable accuracy, this nonuniformity of the sampling must be due to a too short CST simulation time. We checked this issue by the simple MC model for the CST simulation described in Theory and Methods, because here the expected deviations from a uniform sampling can be reliably determined.

The gray bars in Figure 1(b) measure the standard deviations σ_i of the χ_i resulting from 1000 MC model simulations covering the same number of exchange trials as our CST simulation. One would now expect that erf(σ) = 68% of the χ_i are found at smaller deviations than σ_i. In fact, 12 of the 18 χ_i are within the corridor marked by the σ_i, which nicely reproduces the expected statistics. Quite clearly the standard deviations σ_i can be reduced by extending the simulation time, which will then also lead to a CST sampling close to uniformity. Now the question is whether one can estimate the simulation time required for a reasonably uniform sampling. This issue can be addressed by considering the round trip rate τ^{-1} given by eq 17.

CST Round Trip Rates. From the MC model simulations we calculated an average round trip rate τ^{-1} of 0.83 ns^{-1} with a standard deviation of 0.1 ns^{-1}. Equation 17 gives an exact expression for τ applying to the SEO exchange scheme used in the CST simulation. This expression rests on the assumption of identical acceptance probabilities $\bar{P}$ for exchanges along the ladder. The $\bar{P}$ determined from the CST simulation is about 26%, and the resulting value $\tau^{-1} = 0.85$ ns^{-1} is very close to the MC result.

However, the round trip rate observed in the CST simulation is sizably smaller measuring 0.64 ns^{-1}. This deviation suggests that the time interval of 0.5 ps between subsequent exchange trials is too short to yield statistically independent configurations, i.e. that the autocorrelation time of the energy exceeds 0.5 ps. Thus, the system still has some memory of the previous exchange trial, which, however, is tolerable for most practical purposes. Furthermore, a round trip rate of 0.64 ns^{-1} means that only 16 round trips were counted during the CST simulation which is the main cause for the observed 16% RMSD from uniform sampling. To half this RMSD, a 4-fold number of round trips and, thus, a 4-fold simulation time would be necessary. Accordingly, one can estimate the number of round trips needed to achieve a desired level of uniform sampling. In turn, one can *a priori* estimate the required simulation time by multiplying this number by the predicted round trip time given in eq 17.

Table 3. Weights Determined from the REST/A Simulation[a]

i	T_i	w_i (trapezoid)	w_i (WHAM)	Δw_i	$\hat{\chi}_i$
0	300K	0.0	0.0	0.00	1.12
1	350K	−16.91	−16.89	−0.02	1.09
2	415K	−35.64	−35.52	−0.12	0.99
3	500K	−56.16	−55.83	−0.33	0.80

[a] The weights w_i were determined from the initial 100 ps of the REST/A simulation by the trapezoid rule eq 13 and the WHAM formula eq 14 together with the deviations Δw_i and the corresponding uniformity measures $\hat{\chi}_i$. The WHAM weights serve as starting values for the SST/A and SST/B simulation.

Next, we will study the sampling behavior of SST for which we will additionally examine the adaptation scheme based on the WHAM formula.

Reliability of the Trapezoid Rule for SST. Table 3 shows initial weights w_i determined from the short REST/A simulation. Because the solvent−solvent interactions do not contribute to the partition function of SST, these weights are tiny compared to those given in Table 2. Furthermore, the deviations Δw_i between the trapezoid rule eq 13 and the WHAM formula eq 14 are much larger than those listed in Table 2. The associated uniformity measures $\hat{\chi}_i$ predict that errors of this size will lead to a considerable nonuniformity of the SST sampling if the w_i are calculated by the trapezoid rule. Recall here that this rule can be derived based on two assumptions: (i) the heat capacity at consecutive temperatures T_i and T_{i+1} is constant and (ii) the logarithm $\ln(1 + \Delta T_{i+1,i}/T_i)$ is well approximated by its first order Taylor expansion. These conditions are harder to fulfill for SST than for CST because here the temperature steps $\Delta T_{i+1,i}$ are larger.

Figure 2 compares the effects of applying the trapezoid and WHAM rules, respectively, for updating the w_i during SST simulations. Figure 2(a) shows the deviation of the weight w_3 belonging to $T_3 = 500$ K during the SST/A (gray line, trapezoid) and SST/B (black line, WHAM) simulations from the initial value. The first update drastically changes w_3 in both simulations indicating that w_3 has been poorly estimated by the preparatory simulation. The subsequent updates reduce the large initial change to a final deviation of about 1.4 in both cases.

At first glance, the small difference of the resulting w_3 values suggests that the errors of the trapezoid rule are much smaller than predicted by the preparatory REST/A simulation. To check this issue, we have recalculated the weights w_i of the SST/A simulation *a posteriori* by the WHAM formula. The resulting time evolution of w_3 is depicted in Figure 2(a) by the gray dotted line. The difference of 0.22 ± 0.02 between the dotted gray and the solid gray lines is nearly constant during the simulation. Obviously, the trapezoid rule systematically underestimates w_3. A similar underestimate appears already in the initial guess for Δw_3 given in Table 2. Because of this systematic error of the trapezoid rule, the uniformity of the temperature sampling is expected to be suboptimal in SST/A.

Figure 2(b) shows the measured (circles) and predicted (squares) uniformity measures χ_i and $\hat{\chi}_i$ of the SST/A simulation. As indicated by the squares, the average deviations Δw_i between the trapezoid and the WHAM rules predict deviations of up to 13% from uniformity. The measured χ_i
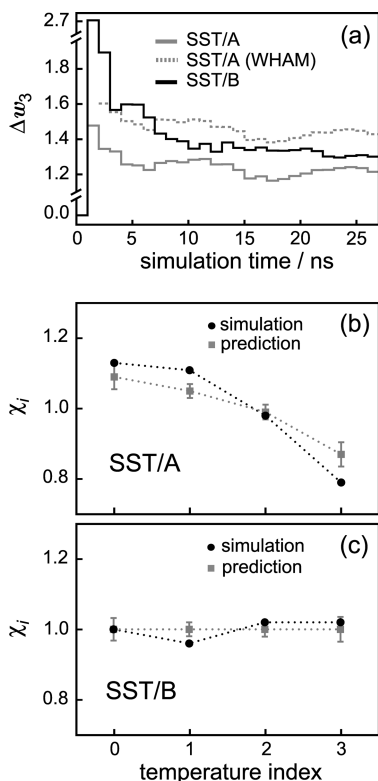
Figure 2. Uniformity of temperature sampling in SST simulations. (a) Deviation of the weight w_3 from its initial value in simulation SST/A (gray) and SST/B (black), respectively. The dotted line shows w_3' calculated a *posteriori* from SST/A using the WHAM expression eq 14. The broken w_3 axis serves to simplify the comparison with Figure 1(a). (b) Measured and predicted uniformity measures χ_i and $\hat{\chi}_i$ of SST/A and (c) of SST/B.

Figure 3. Volumes V_c sampled within 0.25 ns, 0.5 ns, 1 ns, and 2.5 ns by the simulations CST, SST/B, and REST/C. Due to the linearity of the shown $V_c(t)$ curves, the values at 1 ns represent the sampling speeds S in units of n_{lza}/ns. (a) Volumes $V_c(t)$ sampled by the trajectories at all temperatures. According to the inset the $V_c(t)$ curves of REST/C and SST/B are so close that they cannot be distinguished in the main plot. (b) Volume $V_c(t)$ sampled at the target temperature $T_0 = 300$ K. Note that the statistical errors of the measured volumes are smaller than the symbol sizes.

(circles) essentially follow these expectations but show even larger deviations from uniformity, yielding an RMSD of 14%. For instance, χ_3 happens to deviate by about two standard deviations from the respective expectation value $\hat{\chi}_3$. Like for CST, the standard deviations shown as gray bars in the figure were determined from additional MC simulations.

Figure 2(c) compares the uniformity measures of the SST/B simulation. Because the WHAM formula is the reference, the errors Δw_i vanish and eq 16 predicts a uniform sampling $\hat{\chi}_i = 1.0$ at all temperatures. In fact, the measured χ_i are close to 1.0 and show an RSMD of only 3%. Thus, SST/B exhibits an almost perfectly uniform sampling implying that the WHAM formula should be used in SST simulations for updating the w_i. The remaining deviations from uniformity are consistent with the narrow range of the statistical fluctuations estimated by our separate MC simulations. Compared with CST, the much smaller deviations of the χ_i from the predictions $\hat{\chi}_i$ indicate that many more round trips must have occurred during the SST simulations.

SST Round Trip Rates. Equation 17 predicts a round trip rate of 20 ns^{-1} for the simulations SST/(A,B), if the measured average acceptance probability $\bar{P} = 24\%$ is used. Our MC models of SST/(A,B) have reproduced this rate. For the MD simulations SST/A and SST/B, however, we found round trip rates of only 15.8 ns^{-1} and 16.5 ns^{-1}, respectively. Thus, the SST simulations apparently display

the same memory effect which was already observed in the CST simulation and which reduces the round trip rates by 20%. Nevertheless, the SST round trip rates are by a factor of 26 larger than the CST rates. Correspondingly, the χ_i are much better converged in SST than in CST.

The large round trip rates and the nearly uniform sampling achieved in simulation SST/B lead to the expectation that this simulation setting leads for a peptide in solution to an improved statistics. We now will address this issue for our sample peptide 8ALA in TIP3P water.

Sampling Speed of CST, SST, and REST. The purpose of any tempering algorithm is to increase the sampling speed within the conformational space of the studied system. For our various simulation settings we determined the sampling speed using the LZA algorithm described in the Theory and Methods section. This iterative algorithm measures the volume $V_c(t)$ of the configuration space sampled within a simulation time t through an average number n_{lza} of iterations. The simulation speed $S(t)$ is then given by the time derivative of $V_c(t)$.

Figure 3(a) shows the volumes V_c sampled by the simulations CST, SST/B, and REST/C at all temperatures as functions of the simulation time t. The respective sampling speeds S are the constants $V_c(t)/t$ at $t = 1$ ns. The V_c curves of REST/C and SST/B cannot be graphically distinguished at the given scale as is documented by the inset in Figure 3(a). Apparently, CST provides the highest overall sampling speed of the three simulations. The sampling speeds of SST/B and REST/C are by about 10% smaller, which may be caused

Simulated Solute Tempering

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2855**

by the scaling of the solvent part of the Hamiltonian corresponding to an effectively cooler environment.

We have checked the latter conjecture by two MD simulations at 500 K (data not shown) with and without solvent scaling. Here, the effectively cooler solvent indeed reduces the sampling speed of 8ALA by about 20%. For lower temperatures we expect this effect to be correspondingly smaller. However, this small effect is tolerable if the sampling speed at the target temperature $T_0 = 300$ K is sufficiently enhanced, which is after all the aim of solute tempering methods.

Figure 3(b) compares the sampling speeds at $T_0 = 300$ K for the three methods. In contrast to the sampling speed of the generalized ensemble, at 300 K REST/C samples the peptide conformations 2.1 times faster than CST, and SST/B outperforms CST even by a factor of 2.8. Thus, SST/B samples also faster than REST/C, although the two simulations employ the same temperature ladder and the same solvent scaling. An explanation of this speedup is given by the different round trip rates of 11.1 ns^{-1} for REST/C and 16.4 ns^{-1} for SST/B. Due to the higher rate, SST delivers the structural information that is gathered at higher temperatures faster to the target temperature implying an enhanced speed of conformational sampling at T_0.

The reduced round trip rate of REST/C compared to SST/B directly results from the fact that for a given temperature ladder the average acceptance probabilities $\bar{P}_{RE}$ of RE methods (including their sequential versions[13]) are smaller than the probabilities $\bar{P}_{ST}$ of the corresponding ST methods.[40–42] The reason is that in RE the configurations of two replicas must simultaneously meet a certain energy criterion instead of only one replica in ST. Therefore, the average acceptance probability $\bar{P}_{RE}$ should be approximately the square of $\bar{P}_{ST}$. For example, the average acceptance probability of SST/B is about 26%. Thus, we expect a probability of 7% ($0.26^2 \approx 0.07$) for REST/C which is close to the measured value of 9%.

Optimal Exchange Scheme. Because the acceptance probability of REST/C is much smaller than that of SST/B, eq 17 predicts likewise different round trip rates. Compared with that expectation the round trip rate measured for REST/C (11.1 ns^{-1}) seems to be too high compared to SST/B (16.5 ns^{-1}). This large REST/C rate illustrates the advantage of the employed DEO exchange scheme compared to the SEO scheme of SST/B.[52] Furthermore, the optimal exchange probabilities $\bar{P}$ which yield the highest round trip rates are different for these two schemes. For SEO the optimal $\bar{P}$ is 23%,[52,66] whereas for DEO the optimal $\bar{P}$ is between 40% and 45% depending on the ladder size.[39,52]

To investigate the effects of the exchange scheme and the acceptance probability on the SST round trip rates we have carried out the two 12 ns simulations SST/C and SST/D. SST/C switches from SEO to DEO, and SST/D additionally uses five instead of four rungs to span the temperature range from 300 K up to 500 K (see Table 1). SST/D thereby increases $\bar{P}$ to about 40%. For SST/C we found a round trip rate of 20.0 ns^{-1} which is about 20% larger than that of SST/B. Thus, the DEO exchange scheme indeed speeds up the
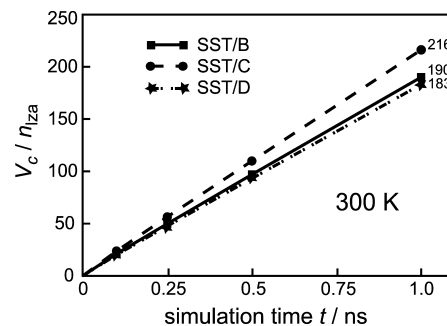

Figure 4. Sampled volumes for the simulations SST/B, SST/C, and SST/D at $T_0 = 300$ K by SST/C. Compared to SST/B the DEO exchange scheme increases the sampling speed by about 15%. This increase corresponds to the enhancement of the round trip rate. Interestingly, the sampling speed of SST/D is smaller than that of SST/B despite the much larger round trip rate. Here, the lower sampling speed is caused by the 25% reduced sampling time at 300 K due to the additional temperature rung, which is not compensated by the higher round trip rate.

Table 4. Round Trip Rates and Sampling Speeds Measured at Temperature $T_0 = 300$ K

label	round trip rates/ns^{-1}	speed/n_{lza}/ns
CST	0.63	68
REST/C	11.1	146
SST/A	15.8	204
SST/B	16.5	190
SST/C	20.0	216
SST/D	23.8	183

round trips sizably. An additional increase of $\bar{P}$ in simulation SST/D leads to a still larger round trip rate of 23.8 ns^{-1}.

Figure 4 shows the effects of the round trip rates, which increase in the sequence SST/B, SST/C, and SST/D, on the sampling speed at 300 K. The highest sampling speed is achieved by SST/C. Compared to SST/B the DEO exchange scheme increases the sampling speed by about 15%. This increase corresponds to the enhancement of the round trip rate. Interestingly, the sampling speed of SST/D is smaller than that of SST/B despite the much larger round trip rate. Here, the lower sampling speed is caused by the 25% reduced sampling time at 300 K due to the additional temperature rung, which is not compensated by the higher round trip rate.

Finally, Table 4 summarizes the round trip rates and sampling speeds measured in our simulations. Using REST instead of CST speeds up the sampling by a factor of 2, whereas SST yields a speedup factor of 3. Note, however, that these factors are conservative estimates because of the particular choice of the 8ALA system. In this system the enthalpic barriers are small, and, therefore, the benefit of tempering methods is limited (see Introduction).[22] Correspondingly, the sampling speeds do not increase very much upon heating the system from the lowest to the highest temperature. However, target applications of tempering methods feature large enthalpic barriers[14,67] for which the sampling has to be accomplished mainly at high temperatures. Correspondingly, the sampling speed at T_0 should depend much stronger on the round trip rates. As a key result,

SST in combination with the DEO exchange scheme should generally show much better sampling properties than REST or CST.

Conclusion

We have introduced simulated solute tempering (SST) which combines the (serial) simulated tempering method with solute tempering, i.e. the key idea of the REST approach. SST poses an efficient alternative to conventional simulated tempering and replica exchange, including REST and its sequential version SREST because it offers the largest acceptance probabilities for a given temperature ladder.

From a practical point of view, it is gratifying to note that SST can be easily implemented. For example, for rigid models of the solvent molecules only the partial charges and the van der Waals parameters have to be scaled to generate the modified Hamiltonians at higher temperatures. Furthermore, SST enables a parallel sampling of many replicas even on heterogeneous computer clusters, because all replicas travel independently through temperature space.

The necessary ingredients of SST are the weights, i.e. the dimensionless free energies of the system at the rungs of the temperature ladder. The trapezoid rule recently suggested by Park and Pande[9] for the computation of the weights is not accurate enough for SST but well suited for CST. Our rederivation of this rule has shown that it is only accurate for temperature ladders featuring small temperature differences, which is the case for CST. In SST a few rungs suffice to span a large temperature range. Due to the failure of the trapezoid rule, the SST weights should be updated using the more complex but asymptotically unbiased WHAM formula of Kumar et al.[45] Then an almost perfectly uniform sampling of the temperature rungs is guaranteed, if the simulation time exceeds the average round trip time by about 2 orders of magnitude.

Our comparison of different sampling methods (REST, CST, SST) applied to an octapeptide in explicit water has demonstrated that the SST sampling is the most efficient one as was shown by the largest round trip rate and the highest sampling speed at T_0. Finally we have shown that the round trip rates can be maximized by using the DEO instead the SEO exchange scheme and by choosing a temperature ladder that provides acceptance probabilities close to 45%.

In conclusion, the sampling efficiency of SST as well as its ease of implementation and application nourishes the hope that simulated tempering will become more popular and that we may see many exciting applications in the future.

Appendix

For a canonical ensemble with a heat capacity C_V independent of the temperature, eq 12 can be derived by the following physical considerations. Using the shorthand notation $\Delta X_{ji} \equiv X_j - X_i$, the entropy difference $\Delta S_{ji} = C_V \ln(1 + \Delta T_{ji}/T_i)$ can be estimated by a first order Taylor expansion of the logarithm as $\Delta S_{ji} \approx C_V \Delta T_{ji}/T_i$. With $C_V = \Delta U_{ji}/\Delta T_{ji}$ one gets

$$\Delta S_{ji} \approx \frac{\Delta U_{ji}}{T_i} \tag{19}$$

With the Helmholtz free energy $F = U - TS$, where U denotes the internal energy, the free energy difference ΔF_{ji} between the systems at T_j and T_i can be written as

$$\Delta F_{ji} = \Delta U_{ji} - T_i \Delta S_{ji} - S_i \Delta T_{ji} - \Delta T_{ji} \Delta S_{ji} \tag{20}$$

Inserting eq 19 one immediately finds

$$F_j \approx F_i - S_i \Delta T_{ji} - \Delta U_{ji} \frac{\Delta T_{ji}}{T_i} \tag{21}$$

With eq 21, the dimensionless free energy difference $\Delta\phi_{ji} = F_j/k_B T_j - F_i/k_B T_i$ can be written as

$$\Delta\phi_{ji} \approx \Delta\beta_{ji}(U_i + \Delta U_{ji}) \tag{22}$$

Interchanging i and j one obtains an equally valid estimate

$$\Delta\phi_{ji} \approx \Delta\beta_{ji}(U_j - \Delta U_{ji}) \tag{23}$$

where we have used $\Delta X_{ij} = -\Delta X_{ji}$. An even better approximation is then given by the arithmetic mean

$$\Delta\phi_{ji} \approx \Delta\beta_{ji} \frac{U_i + U_j}{2} \tag{24}$$

where ΔU_{ji} cancels. Restricting the internal energy U to its configurational part, i.e., to the average potential energy $\langle E \rangle$, yields the "trapezoid" rule eq 12.

Supporting Information Available: Average acceptance probabilities $\bar{P}_{i,i \pm 1}$ for the various simulations (Tables S1 and S2). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers (Peptide Sci.)* **2001**, *60*, 96.

(2) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425.

(3) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.

(4) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.

(5) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.

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

(7) Lyubartsev, A. P.; Martinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776.

(8) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451.

(9) Park, S.; Pande, V. S. *Phys. Rev. E* **2007**, *76*, 016703.

(10) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042.

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

(12) Affentranger, R.; Tavernelli, I. *J. Chem. Theory Comput.* **2006**, *2*, 217.

Simulated Solute Tempering

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2857**

(13) Hagen, M.; Kim, B.; Liu, P.; Friesener, R. A.; Berne, B. J. *J. Phys. Chem. B* **2007**, *111*, 1416.

(14) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749.

(15) Kubitzki, M. B.; de Groot, B. L. *Biophys. J.* **2007**, *92*, 4262.

(16) Xu, W.; Lai, T.; Yang, Y.; Mu, Y. *J. Chem. Phys.* **2008**, *128*, 175105.

(17) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. *Phys. Rev. Lett.* **2006**, *96*, 028105.

(18) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035.

(19) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123*, 154105.

(20) Rick, S. W. *J. Chem. Theory Comput.* **2006**, *2*, 939.

(21) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126*, 014903.

(22) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2*, 1200.

(23) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458*, 244.

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

(25) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46*, 225.

(26) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587.

(27) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. *J. Chem. Phys.* **2004**, *121*, 10748.

(28) Jas, G. S.; Kuczera, K. *Biophys. J.* **2004**, *87*, 3786.

(29) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *J. Mol. Biol.* **2004**, *336*, 241.

(30) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C.-K.; Li, M. A. *Proteins* **2005**, *61*, 795.

(31) Villa, A.; Stock, G. *J. Chem. Theory Comput.* **2006**, *2*, 1228.

(32) Schrader, T. E.; Schreier, W. J.; Cordes, T.; Koller, F. O.; Babitzki, G.; Denschlag, R.; Renner, C.; Lweneck, M.; Dong, S.-L.; Moroder, L.; Tavan, P.; Zinth, W. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15729.

(33) Villa, A.; Widjajakusuma, E.; Stock, G. *J. Phys. Chem.* **2008**, *112*, 134.

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

(35) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386.

(36) Calvo, F. *J. Chem. Phys.* **2005**, *123*, 124106.

(37) Brenner, P.; Sweet, C. R.; VonHandorf, D.; Izaguirre, J. A. *J. Chem. Phys.* **2007**, *126*, 074103.

(38) Sindhikara, D.; Meng, Y.; Roitberg, A. E. *J. Chem. Phys.* **2008**, *128*, 024103.

(39) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2009**, *473*, 193.

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

(41) Park, S. *Phys. Rev. E* **2008**, *77*, 016709.

(42) Zhang, C.; Ma, J. *J. Chem. Phys.* **2008**, *129*, 134112.

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

(44) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435.

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

(46) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.

(47) Kobrak, M. N. *J. Comput. Chem.* **2003**, *24*, 1437.

(48) Berg, B. A. *J. Stat. Phys.* **1996**, *82*, 323.

(49) Bartels, C.; Karplus, M. *J. Comput. Chem.* **1997**, *18*, 1450.

(50) Park, S.; Ensign, D. L.; Pande, V. S. *Phys. Rev. E* **2006**, *74*, 066703.

(51) Zhang, C.; Ma, J. *Phys. Rev. E* **2007**, *76*, 036708.

(52) Lingenheil, M.; Denschlag, R.; Mathias, G.; Tavan, P. *Chem. Phys. Lett.* **2009**. in press (doi:10.1016/j.cplett.2009.07.039).

(53) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435.

(54) Gardiner, C. W. *Handbook of Stochastic Methods*, 2nd ed.; Springer, Berlin, 1985.

(55) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.

(56) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586.

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

(58) Schaftenaar, G.; Noordik, J. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123.

(59) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847.

(60) Niedermeier, C.; Tavan, P. *J. Chem. Phys.* **1994**, *101*, 734.

(61) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Oxford University Press: Oxford, 1987.

(62) Eichinger, M.; Grubmüller, H.; Heller, H.; Tavan, P. *J. Comput. Chem.* **1997**, *18*, 1729.

(63) Kraeutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Comput. Chem.* **2001**, *22*, 501.

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

(65) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, *91*, 164.

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

(67) Reichold, R.; Fierz, B.; Kiefhaber, T.; Tavan, P. Submitted for publication.

JCTC Journal of Chemical Theory and Computation

Adsorption and Diffusion of Alkanes in Na-MOR: Modeling the Effect of the Aluminum Distribution

Shuai Ban[†] and Thijs J. H. Vlugt*[,‡]

Condensed Matter and Interfaces, Department of Chemistry, Utrecht University, P.O. Box 80.000 3508TA Utrecht, The Netherlands and Delft University of Technology, Process & Energy Laboratory, Leeghwaterstraat 44, 2628CA Delft, The Netherlands

Abstract: We investigated the adsorption and the diffusion of alkanes in the sodium-exchanged zeolite Mordenite (Na-MOR) using molecular simulations. MOR-type zeolite consists of main channels (6.5 × 7 Å) oriented along the *z* crystallographic axis that are connected to small side pockets (3.4 × 4.8 Å). It is well-known that the adsorption of alkanes in Na-MOR strongly depends on the precise location of the framework Al atoms either in the main channel or the side pockets (Calero and co-workers, *Angew. Chem. Int. Ed.* **2007,** *46,* 276). We found that this effect can be characterized by a single-order parameter: the number of framework Al in the main channel divided by the number of framework Al in the side pocket (M/S ratio). For any M/S ratio, the adsorption isotherm follows from a linear interpolation between the reference isotherms. This enabled us to predict adsorption isotherms for any distribution of the Al framework atoms or estimate the M/S ratio for a given isotherm. We found that the same model can predict the effect of the M/S ratio on the self-diffusion coefficient, the Maxwell−Stefan diffusion coefficient, and the accessible micropore volume.

1. Introduction

Mordenite (MOR-type zeolite) is an important catalyst in the petrochemical industry, as it is used for the acid-catalyzed isomerization of alkanes and aromatics.[1,2] For example, Pt/H-MOR is a suitable catalyst for the hydro-isomerization of linear alkanes to branched ones, even though slow intracrystalline transport hinders the catalytic performance to some extent.[3] Therefore, it is important to understand the adsorption and transport of hydrocarbon molecules in this zeolite.[4] The framework of MOR-type zeolite consists of main channels formed by large 12-membered rings with an elliptical shape of 6.7 × 7.0 Å oriented along the *z* crystallographic axis. Small side pockets are connected to these main channels by eight-membered rings (3.4 × 4.8 Å) that are oriented along the *y* crystallographic axis.[5] It was found that methane is the only hydrocarbon that can be adsorbed in the side pockets.[6,7]

In general, the structure of zeolites consists of covalently bonded TO_4 units, in which the framework T atom is usually a silicon (Si) or an aluminum (Al) atom. To obey charge neutrality, the substitution of a Si atom by an Al atom requires the presence of a nonframework cation (usually Na^+ or K^+) or a proton (H^+). It is well-known that the adsorption of guest molecules in Na-MOR is significantly enhanced by nonframework Na^+ cations.[8] The positions of the nonframework cations in MOR-type zeolite are coupled to the positions of the framework Al atoms.[9] The *average* distribution of framework Al over the various T sites (labeled T_1, ..., T_4) in Na-MOR was obtained experimentally by detecting Brønsted acid sites.[9−12] It was found that the majority of Al is located at the T_3 and T_4 sites close to the side pockets. Knowledge on the distribution of Al among T sites alone is not sufficient to obtain the *exact* individual positions of all framework Al atoms. Recently, Calero and co-workers used an elegant reversed engineering approach to identify the positions of the framework Al atoms in zeolites.[13,14] For a fixed Si/Al ratio, this approach considers all possible

---

* Corresponding author. E-mail: t.j.h.vlugt@tudelft.nl.

† Utrecht University.

‡ Delft University of Technology.

Alkanes in Na-MOR

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2859**

distributions of Al in the unit cell of the zeolite. It was found that this approach correctly predicts the Al distribution in Na-MOR. These authors also found that for some zeolites, e.g., LTA, FAU, and MFI, the precise positions of the framework Al atoms do not influence adsorption, while for other zeolites, e.g., MOR, FER, and TON, the adsorption strongly depends on the positions of the framework Al atoms. For MOR-type zeolite, this is because nonframework Na^+ cations in the side pocket only have a weak interaction with adsorbed guest molecules in the main channel, while this interaction is much stronger for cations located in the main channel.

Many experimental and simulation studies have investigated the effect of nonframework cations on the diffusivity of guest molecules, particularly in MFI-type zeolite.[15−18] Fan et al.[16] used molecular dynamics simulations to study the diffusion of ethene in MFI-type zeolite. These authors found that the diffusion of ethene was slowed down from 2.7 × 10^{-9} m²/s in a silicalite framework to 1.6 × 10^{-9} m²/s in ZSM-5 with four Al per one unit cell. Masuda et al.[18] studied the influence of acid sites on the diffusion of aromatics in MFI-type zeolite. This study revealed that at low temperatures, the diffusivity significantly decreased with an increasing number of acid sites, while the diffusion of aromatics was hardly affected by the acid sites at high temperatures. Leroy et al.[15] combined the quasi-elastic neutron-scattering technique with molecular simulations to study the diffusion of alkanes in MFI. These authors found that the diffusivity of octane in silicalite is four times larger than in Na-ZSM-5. This result agreed well with their molecular dynamics simulations for MFI-type zeolite with two Na^+ per unit cell. In summary, previous studies show that diffusion of hydrocarbons is slowed down by nonframework cations. Therefore, it is important to know the effect of the Al distribution on this.

In this work, computer simulations were used to investigate the influence of the positions of framework Al on the adsorption and diffusion properties of Na-MOR in a more coarse-grained way. In our simulations, the Al distribution was described by a single-order parameter: the M/S ratio, which is the ratio of the number of framework Al in the main channels (M) and the side-pockets (S) in MOR-type framework (M/S ratio). The reasons for using this order parameter were: (1) The adsorption of guest molecules is only sensitive to the framework Al located in either the main channels or the side pockets, rather than the exact positions of framework Al atoms. (2) The Al distribution is a more accessible parameter for experiments than the distribution of nonframework Na^+.

This paper is organized as follows. In Section 2, we briefly describe the construction of a Na-MOR supercell as well as the simulation methods for computing the adsorption isotherms, the diffusivities, and the accessible micropore volume. In Section 3.1, the computed Na^+ distribution in MOR is compared with experiments. In Section 3.2, we present an interpolation model to predict isotherms in Na-MOR for any M/S ratio, using two reference isotherms with a known M/S ratio. In Section 3.3, we show that the nonframework Na^+ cations seriously hinder the diffusion of
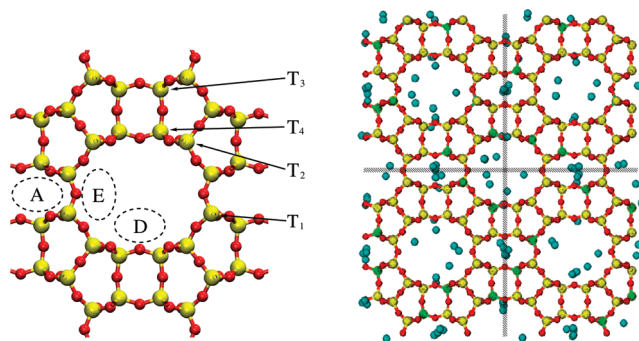

Figure 1. (a) The preferential sites A, D, and E for nonframework Na^+ cations in Na-MOR. The colors are: red for O and yellow for Si and Al. The four T sites T_1, T_2, T_3, and T_4 are shown as well. (b) Typical snapshot of Na-MOR with Si/Al = 5 for a supercell consisting of 2 × 2 × 4 unit cells at 300 K. The colors are: red for O, cyan for Na^+, yellow for Si, and black for Al.

alkanes in Na-MOR at low temperature, and this effect is correctly described by our model. We show that our model is also applicable for calculating the micropore volume of Na-MOR (Section 3.4). Our findings are summarized in Section 4.

2. Simulation Methods

2.1. The Na-MOR Lattice. Zeolite Na-MOR consists of Na, Si, Al, and O atoms with the composition $Na_xAl_xSi_{48-x}O_{96}$. Natural mordenite has a Si/Al ratio of approximately 4.3−6.0, so $x \approx 8$.[10] Demuth et al.[19] pointed out that the most common space group for MOR-type zeolite is *Cmcm* with the exception of dehydrated protonated MOR which has space group *Pbcn*. Macedonia et al.[20] showed that differences in adsorption isotherms between the space groups *Cmcm* and *Pbcn* are negligible. Therefore, in this paper we restricted ourselves to the Na-MOR lattice with a *Cmcm* symmetry. The dimensions of the MOR unit cell are 18.094 × 20.516 × 7.524 Å with $\alpha = \beta = \gamma = 90°$. Each unit cell contains four different tetrahedral sites (T_1, ..., T_4) for Si and Al framework atoms, and ten different oxygen sites (O_1, ..., O_{10}).[21] The T_3 sites are located in the side pocket, while T_1, T_2, and T_4 are located in the main channel, see Figure 1a. Alberti et al.[10] derived the Al distribution over the four T sites in natural Na-MOR from X-ray diffraction (XRD) measurements. This is also expected to be applicable to synthetic Na-MOR and H-MOR.

In this work, we constructed large supercells of MOR-type zeolite consisting of 2 × 2 × 4 unit cells. Starting from an all-silica supercell, Si framework atoms are randomly substituted by Al in such a way that (1) the Löwenstein rule is obeyed, and (2) the relative Al content of each T site corresponds to the given M/S ratio. As a result, many Na-MOR supercells could be generated with the same Si/Al ratio but with different positions of the framework Al. We will show that the properties of these supercells are nearly identical, as the relative Al content of the T sites in the main channel and the side pocket is identical for these supercells. Computed properties were then averaged over 5−10 supercells.

2.2. Computing Adsorption Isotherms. Adsorption isotherms were computed by using configurational-bias Mon-

te Carlo[22−24] simulations in grand-canonical ensemble (GCMC).[24−27] Interactions between the guest molecules, the zeolite framework, and the nonframework cations were described by Lennard-Jones interactions. In addition, electrostatic interactions between the nonframework Na$^+$ cations and the framework were included. All force field parameters were taken from the work of Calero et al.[28] The Ewald summation was used to compute electrostatic interactions. A typical simulation consisted of 3×10^6 Monte Carlo cycles. In each cycle, trial moves were chosen at random with a fixed probability: translation of a guest molecule or a nonframework cation (15%), rotation of a guest molecule (15%), exchange of a guest molecule with a particle reservoir (55%), and partial regrowth of a guest molecule (15%). The number of trial moves in each cycle was equal to the number of adsorbed guest molecules with a minimum of 20. The number of Na$^+$ atoms in Na-MOR was constant during the simulations. For the simulation of adsorption, the zeolite framework was kept rigid, and the nonframework cations were mobile. For the systems considered here, including framework flexibility usually results in a negligible deviation of adsorption isotherms.[29] For more details on these simulations, we refer the reader to refs 26−28.

2.3. Computing Diffusivities of Guest Molecules. Molecular dynamics simulations[30] were used to calculate the self- and Maxwell−Stefan diffusion coefficients of adsorbed guest molecules in MOR-type zeolite. We used the velocity Verlet algorithm[31] with a time step of 0.5 fs. The temperature was controlled by a Nosé−Hoover thermostat.[24] The initial configuration of guest molecules was taken from the final configuration of a Monte Carlo simulation in the NVT ensemble. To avoid single-file diffusion, each main channel of MOR-type zeolite contained only a single guest molecule. For simplicity, we used a rigid framework. Nonframework cations were allowed to move freely in the zeolite. The self- (D^{self}) and Maxwell−Stefan diffusion coefficients (D^{MS}) of a single component adsorbed in a zeolite were computed from particle displacements:

$$D_\alpha^{self} = \frac{1}{2n} \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \left\langle \left(\sum_{i=1}^n (r_{i,\alpha}(t + \Delta t) - r_{i,\alpha}(t))^2 \right) \right\rangle$$

(1)

$$D_\alpha^{MS} = \frac{1}{2n} \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \left\langle \left(\sum_{i=1}^n (r_{i,\alpha}(t + \Delta t) - r_{i,\alpha}(t)) \right)^2 \right\rangle$$

(2)

in which α is the direction of diffusion, n is the number of adsorbed guest molecules, and $r_i(t)$ is the position of molecule i at time t.[32−34] For MOR-type zeolite, only diffusion in the main channel (z direction) was taken into account, i.e., $\alpha = z$. In practice, the mean-square displacement in eqs 1 and 2 was computed using an order-n algorithm.[24,33,35]

2.4. Pore Size Distribution. In this work, the pore size of a certain cavity or channel was defined as the maximum diameter of a sphere that can be located in there. This definition is applicable to pores with an arbitrary shape. For cylindrical or slit pores, our definition is identical to the IUPAC definition.[36] Using this definition, the pore size

Table 1. Comparison of the Na$^+$ Occupancies between Experiments and Simulations

	Si/Al	site A, %	site D, %	site E, %
experiment[a]	5.7	43	36	21
experiment[b]	5.7	53	34	13
experiment[c]	5	49	32	19
simulation[d]	5	50	38	12
this work	5.0	40.2	35.5	20.9
this work	5.5	40.1	35.3	20.9

[a] Schlenker et al.[5] [b] Devautour et al.[41] [c] Coughlan et al.[39,44] [d] Tyburce et al.[45] The Na$^+$ occupancy was averaged over five supercells. Note that the differences in occupancy between the supercells were very small.

distribution was computed from the coordinates of the framework atoms using the following algorithm:[37] (1) A three-dimensional grid with a small spacing is constructed. We typically use a grid size of 0.1 Å (in each direction). (2) A spherical test particle is positioned at a random position in the zeolite. The radius of this particle is chosen as the minimum distance between the center of the particle and any of the zeolite framework atoms minus the radius of the closest framework atom. The radius of an oxygen framework atom is 1.35 Å, and the radius is 0.99 Å for Na$^+$.[21] Tetrahedral atoms, e.g. Si, Al, are not considered, as for most zeolites they are well screened by adjacent oxygen atoms.[38] (3) The diameter of the test particle is recorded for all grid points that are inside the spherical particle. (4) This procedure is repeated many times. We found that the number of test spheres should be at least 100 times the number of grid points. For each grid point, the maximum recorded diameter is computed, and this quantity is defined as the local pore size of a specific grid point. The simulation stops when the local pore size of all grid points is converged. (5) The fraction of pores with a diameter between r and $r + \Delta r$ (pore size distribution) is equal to the fraction of grid points with a maximum diameter between r and $r + \Delta r$. Using our calculated pore size distribution, the micropore volume was calculated by integrating the pore size distribution from 4.5 to 20 Å. For more details, we refer the reader to ref 37.

3. Results and Discussion

3.1. Na$^+$ Distribution. Nonframework Na$^+$ cations interact with the framework Al through Lennard-Jones and long-range Coulombic interactions. As a consequence, Na$^+$ cations are not directly bonded to framework Al atoms. Experiments on dehydrated Na-MOR crystals[5,39−41] identified three favorable locations for nonframework Na$^+$: the center of eight-membered rings that run parallel to the main channels (site A), the main channel at the entrance of the side pocket (site D), and the main channel far away from the side pocket (site E), see Figure 1a. Our simulations showed the same adsorption sites for Na nonframework cations, see Figure 1b. The Na$^+$ occupancies at these sites have been computed for natural Na-MOR with Si/Al = 5 and 5.5, respectively. Table 1 shows that the computed Na$^+$ occupancies at these sites agree very well with the available experiments. For different supercells with an identical distribution of Al over the T sites, we found that the occupancy of Na$^+$ at the three sites is nearly identical. The
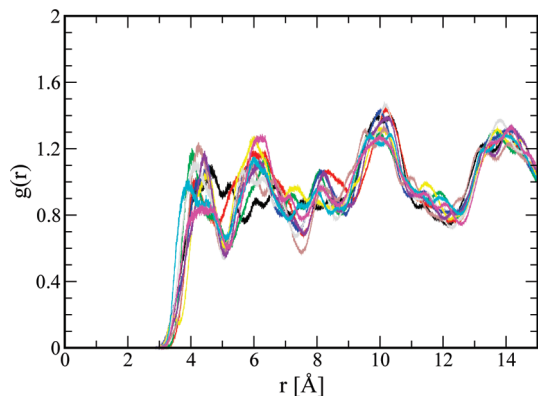
Alkanes in Na-MOR

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2861**

Figure 2. Radial distribution functions for Na$^+$−Na$^+$ interactions for 10 supercells with the same Al distribution (Si/Al = 5), see Section 2.1.

corresponding radial distribution functions for Na$^+$−Na$^+$ interactions are very similar, as they all show peaks at distances of 4.3, 6.2, 10.1, and 14 Å, see Figure 2.

3.2. Prediction of Adsorption Isotherms. Na-MOR has two main adsorption sites for alkanes, the large main channels and the smaller side pockets. Methane is the only alkane that can access the side pockets.[6,7] The side pockets are also accessible for nonframework Na$^+$ cations. It is natural to divide the nonframework Na$^+$ cations into two groups: Na$^+$ in the main channels (denoted by Na$_M^+$) and Na$^+$ in the side pockets (denoted by Na$_S^+$). Na$_M^+$ corresponds to the Na$^+$ cations at site D and E, while Na$_S^+$ corresponds to Na$^+$ at site A. The Na$_M^+$ cations interact more strongly with alkanes absorbed in the main channel, as they are more close by. A convenient order parameter to characterize the Na$^+$ distribution is the Al M/S ratio, which is the ratio of the number of the framework Al in the main channels and the side pockets. The reason for this is that, due to the strong interactions between Al and Na$^+$, the distribution of the Na$^+$ cations approximately follows that of the framework Al. The difference of the adsorption properties between different Na-MOR supercells with the same Al M/S ratio was found to be small in our simulations. Al framework atoms at the crystallographic site T$_3$ are considered to be in the side pocket, while Al atoms at T$_1$, T$_2$, and T$_4$ are considered to be in the main channel, see Figure 1a. The same classification was previously used by IR spectroscopy experiments of probe molecules in MOR-type zeolite.[42] This technique is used to locate Brønsted acid sites in H-MOR. We expect that the Al distribution in Na-MOR is the same as in H-MOR, as H-MOR is usually obtained by exchanging Na$^+$ with H$^+$.[43]

To model the adsorption behavior of alkanes in Na-MOR, as a function of the M/S ratio (Al content in the main channel divided by the Al content in the side pocket, here denoted by r), we constructed the following model. The total number of Al atoms per unit cell is denoted by n, while n_M and n_S are the number of Al atoms per unit cell in the main channels and the side pockets, respectively. Of course:

$$n = n_M + n_S \tag{3}$$

and we define the ratio r by

$$r = \frac{n_M}{n_S} \tag{4}$$

From this it directly follows that:

$$n_M = n\frac{r}{r+1} \tag{5}$$

$$n_S = n\frac{1}{r+1} \tag{6}$$

Our key assumption is that the loading of guest molecules at pressure P (here denoted by $I(P)$) is a linear function of n_M and n_S:

$$
\begin{aligned}
I(P) &= I_0(P) + \alpha_M(P)n_M + \alpha_S(P)n_S \\
&= I_0(P) + \frac{n}{r+1}[\alpha_M(P)r + \alpha_S(P)]
\end{aligned} \tag{7}
$$

where $\alpha_M(P)$ and $\alpha_S(P)$ are pressure-dependent constants and $I_0(P)$ is the loading of guest molecules for all-silica MOR at pressure P. The maximum number of Al at the T$_3$ sites is four Al atoms per unit cell ($n_S \leq 4$), which is much lower than the maximum possible number of Al in the main channel. For a given total content of framework Al atoms, there are two extreme situations: (1) the M structure, i.e., all Al is located in main channels ($n_S = 0$, $n_M = n$, the corresponding adsorption isotherm is denoted by $I_M(P)$) and (2) the S structure, where the side pockets are fully loaded with Al, and the remaining Al are located in the main channel (adsorption isotherm denoted by $I_S(P)$). This leads to

$$I_M(P) = I_0(P) + \alpha_M(P)n \tag{8}$$

and

$$I_S(P) = \begin{cases} I_0(P) + \alpha_S(P)n & \text{for } n \leq 4 \\ I_0(P) + 4\alpha_S(P) + (n-4)\alpha_M(P) & \text{for } n > 4 \end{cases} \tag{9}$$

By eliminating $\alpha_M(P)$ and $\alpha_S(P)$ we obtain

$$I(P) = \begin{cases} \dfrac{r}{r+1}I_M(P) + \dfrac{1}{r+1}I_S(P) & \text{for } n \leq 4 \\ \dfrac{4r-n+4}{4(r+1)}I_M(P) + \dfrac{n}{4(r+1)}I_S(P) & \text{for } n > 4 \end{cases} \tag{10}$$

When the Si/Al ratio is larger than 11 (i.e., $n \leq 4$), the weight factors of $I_M(P)$ and $I_S(P)$ in eq 10 are simply the fractions of framework Al at the sites M and S.

To test the prediction of eq 10, we computed adsorption isotherms of propane and butane in Na-MOR. These molecules are exclusively adsorbed in the main channels. Figure 3 shows the computed adsorption isotherms I_M and I_S for Na-MOR with four Al per unit cell. Both isotherms significantly differ from the one for an all-silica structure. As expected, the adsorption isotherm strongly depends on the positions of the framework Al atoms, and the differences between isotherms of different supercells are small. At a given pressure, the adsorbed amount for the S structure is always lower than that of the M structure due to the increased average distance between the guest molecules and the nonframework Na$^+$ cations.
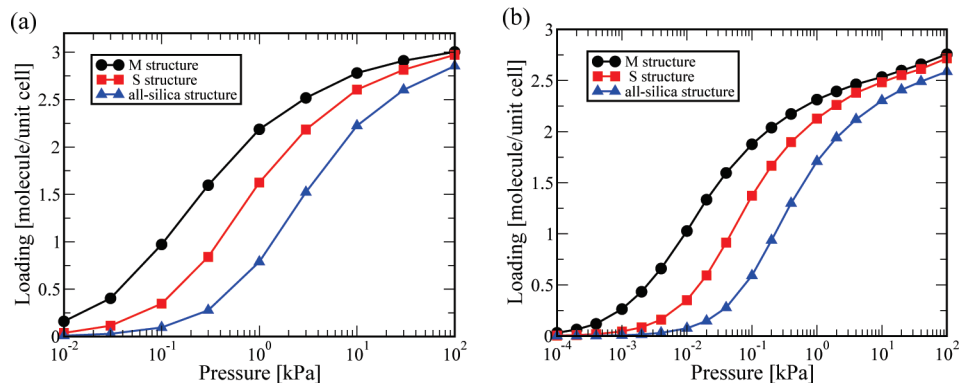
Figure 3. Adsorption isotherms of propane (a) and butane (b) in Na-MOR at 300 K for Si/Al = 11. The M and S structures are shown as well as the isotherm for the all-silica structure. Each isotherm is the average over three supercells with the same M/S ratio. The error bars indicate the differences between these supercells. For each supercell, the error in the computed loading is much smaller than the symbol size.
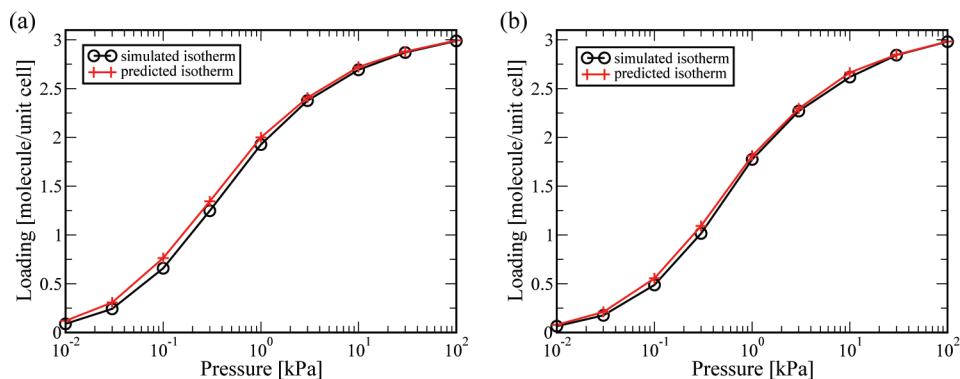

Figure 4. Linear interpolation of adsorption isotherms for propane in Na-MOR with Si/Al = 11 at 300 K. The predicted isotherm using eq 9 was compared to the computed isotherm at the same Al M/S ratio. (a) Al M/S ratio 2:1 and (b) Al M/S ratio 1:2.
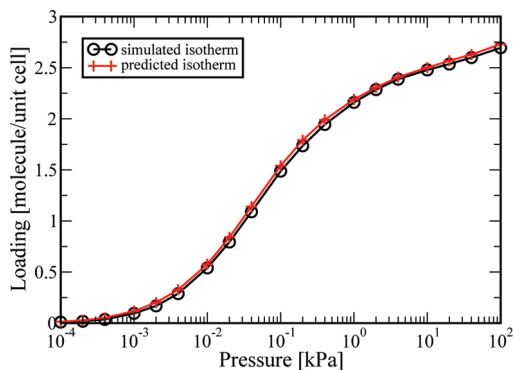
Figure 5. Linear interpolation of adsorption isotherms for butane in Na-MOR with Si/Al = 11 at 300 K. The predicted isotherm using eq 10 was compared to the computed isotherm at the same Al M/S ratio 1:2.

Figure 6. Comparison of the predicted isotherms (eq 10) with experimental data[7] for propane and butane in Na-MOR with Si/Al = 19.75 at 373 and 438 K. The estimated Al M/S ratio was 3:1 for this system.

Figures 4 and 5 show that the prediction of eq 10 is excellent. This result is useful in two ways: First, the adsorption isotherm for an arbitrary M/S ratio can be predicted by linear interpolation between two reference isotherms (for which we know the M/S ratio). Second, the M/S ratio in a Na-MOR sample can be estimated by fitting a measured isotherm to two reference isotherms. As an example, we estimated an M/S ratio of 3:1 from experimental adsorption isotherms of propane and butane in Figure 6. This estimated ratio is close to the ratio of 2:1 proposed by Alberti et al.[10]
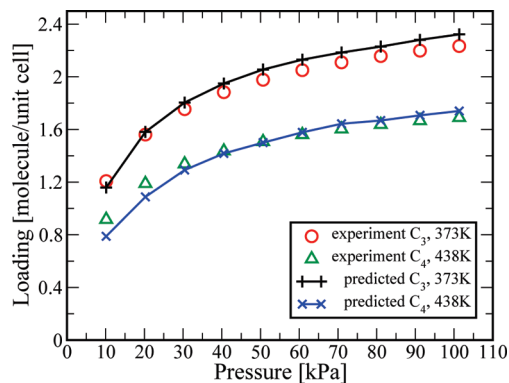
In Figure 7, it is shown that our interpolation model is able to accurately predict the Henry coefficients of *n*-alkanes with 2−5 carbon atoms. For longer *n*-alkanes, small differences appeared between the computed and the predicted Henry coefficients. The reason for this is that longer chains will have interactions with more Na$^+$ cations at the same time, leading to nonlinear effects.

3.3. Prediction of Diffusivities. The computed and predicted diffusivities of alkanes in Na-MOR for various Si/Al ratios are shown in Table 2. The simulations were performed for propane and hexane diffusing in the all-silica MOR and
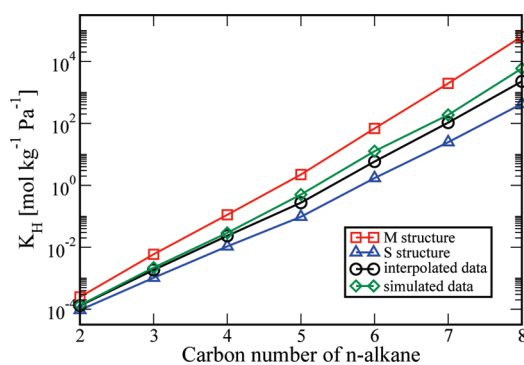
Figure 7. Linear interpolation of the Henry coefficients of *n*-alkanes in Na-MOR with Si/Al = 11 at 300 K. The error bars are much smaller than the symbols.

Table 2. Self- and Maxwell−Stefan Diffusion Coefficients of Propane and Hexane in All-Silica MOR and Na-MOR with Si/Al = 5:23 at 300 K and 600 K[a]

	Si/Al	C_3 at 300 K	C_6 at 300 K	C_3 at 600 K	C_6 at 600 K
D^{self}	∞	5.5	7.0	12.3	13.1
D^{self}	23	2.5	1.6	7.0	5.7
D^{self}	5	1.8	0.9	5.5	3.6
D^{MS}	∞	5.5	7.0	11.6	12.8
D^{MS}	23	2.3	1.4	6.9	5.5
D^{MS}	5	1.8	0.9	5.3	3.5

[a] Each main channel contained a single alkane molecule. Self- and Maxwell−Stefan diffusivities are denoted by D^{self} and D^{MS} in units of 10^{-8} m²/s, respectively. The diffusion coefficients in Na-MOR were obtained by averaging over 10 supercells with the same M/S ratio. The differences between supercells are approximately 15%. For each supercell, the error in the computed diffusivity was around 20% for D^{MS} and 10% for D^{self}. For all-silica MOR, the error in the computed diffusion coefficients was smaller than 1%.

Na-MOR containing two or eight Na⁺ cations at 300K and 600K, respectively. As each main channel contains a single alkane molecule only, it is expected that the Maxwell−Stefan and the self-diffusivities do not differ very much. It can be seen from Table 2 that both the self- and Maxwell−Stefan diffusivities decrease as a function of the Na⁺ concentration irrespective of the temperature and the alkane chain length. This is in line with the results from Leroy et al.[15] Note that the differences of the calculated diffusivities were about 15% for supercells with the same Al distribution. For each supercell, the simulation error is about 20% for D^{MS} and 10% for D^{self}. These errors are much smaller than the difference of the diffusivities for Na-MOR with different Si/Al ratios. Therefore, we can conclude that the diffusion of alkanes is slowed down by the nonframework cations of Na-MOR.

The self- and Maxwell-Stefan diffusivities were computed for propane in Na-MOR with Si/Al = 11 at 300 K and 600 K at a loading of one propane per main channel, see Table 3. As the loading is quite low, the self- and Maxwell−Stefan diffusivities are almost equal. However, the diffusivity of propane in the S structure is more than twice as large as in the M structure. The reason for this is that because of the strong interactions between propane and Na⁺, the Na⁺ ions in the main channels will slow down the diffusion of propane significantly. Therefore, $D_M < D_{M/S = 2:1} < D_{M/S = 1:2} < D_S$. In the same spirit as eq 10, we predicted the self- and

Table 3. Interpolation of Propane and Hexane Diffusivities in Na-MOR with Si/Al = 11 at 300 K and 600 K, Respectively[a]

		D_M	D_S	$D_{M/S = 1:2}$	$D_{M/S = 2:1}$
		C_3 (300 K)			
simulated	D^{self}	1.4	2.8	2.0	1.6
predicted	D^{self}	n.a.	n.a.	2.3	1.9
simulated	D^{MS}	1.3	3.0	2.2	1.4
predicted	D^{MS}	n.a.	n.a.	2.4	1.9
		C_6 (600 K)			
simulated	D^{self}	3.1	5.1	4.5	3.7
predicted	D^{self}	n.a.	n.a.	4.4	3.8
simulated	D^{MS}	2.9	5.1	4.6	3.5
predicted	D^{MS}	n.a.	n.a.	4.4	3.6

[a] Each main channel contained a single alkane molecule. Self- and Maxwell−Stefan diffusivities were denoted by D^{self} and D^{MS} in units of 10^{-8} m²/s. D_M and D_S are the computed diffusivities for the M and S structure, see Section 3.2. All reported values were obtained by averaging over 10 supercells with the same M/S ratio. The differences between supercells are approximately 12%. For each supercell, the error in the estimated diffusivity was approximately 20% for D^{MS} and 5% for D^{self}.

Table 4. Interpolation of the Micropore Volume (*V*) of Na-MOR with Si/Al = 11 in Units of mL/g[a]

	V_M	V_S	$V_{M/S = 1:2}$	$V_{M/S = 2:1}$
simulated	0.1816	0.1934	0.1874	0.1855
predicted	n.a.	n.a.	0.1894	0.1855

[a] The notation is the same as in Table 3. All values were obtained by averaging over 10 supercells with the same Al M/S ratio. The differences of the micropore volumes were 0.5% for the 10 supercells with the same Al M/S ratio. The simulation error of the micropore volume for each supercell was approximately 0.2%.

Maxwell−Stefan diffusivities for M/S = 2:1 and M/S = 1:2, based on the computed diffusivities D_M and D_S:

$$D = \begin{cases} \dfrac{r}{r+1}D_M + \dfrac{1}{r+1}D_S & \text{for} \quad n \leq 4 \\[2ex] \dfrac{4r - n + 4}{4(r+1)}D_M + \dfrac{n}{4(r+1)}D_S & \text{for} \quad n > 4 \end{cases}$$

(11)

In this equation, *D* is the predicted diffusivity and D_M and D_S are the two reference diffusivities. The predicted diffusivities are very close to the computed ones. Differences of calculated diffusivities are about 12% for the supercells with the same Al M/S ratio. For each supercell, the simulation error is about 20% for D^{MS} and 5% for D^{self}. These errors are much smaller than the difference of the diffusivities for Na-MOR with different Al M/S ratios. Therefore, we can conclude that the interpolation scheme is able to predict self- and Maxwell−Stefan diffusivities in Na-MOR.

3.4. Prediction of the Micropore Volume. The micropore volume of Na-MOR for different M/S ratios is shown in Table 4. As the micropore volume was calculated from pores in the range of 4.5−20 Å, the main contribution to the micropore volume originates from the main channels. Therefore, when all the nonframework Na⁺ is located in the main channel, the micropore volume decreases. This leads to the following prediction: $V_M < V_{M/S = 2:1} < V_{M/S = 1:2} < V_S$. For a given M/S ratio, the micropore volume can be computed using

$$V = \begin{cases} \dfrac{r}{r+1}V_{\mathrm{M}} + \dfrac{1}{r+1}V_{\mathrm{S}} & \text{for} \quad n \leq 4 \\ \dfrac{4r-n+4}{4(r+1)}V_{\mathrm{M}} + \dfrac{n}{4(r+1)}V_{\mathrm{S}} & \text{for} \quad n > 4 \end{cases}$$

$$(12)$$

Table 4 shows that the predicted micropore volumes are in good agreement with the computed ones. The differences in micropore volume are 0.5% for supercells with the same M/S ratio. The error in the computed micropore volume for each supercell is about 0.2%. These errors are more than 1 order of magnitude smaller than the difference in the micropore volume between the M and S structure. Therefore, we can conclude that the interpolation scheme works also well for the prediction of the accessible micropore volume in Na-MOR.

4. Conclusions

We presented a simple interpolation model to predict the effect of the framework Al distribution on the adsorption isotherms and diffusivities of alkanes in Na-MOR as well as the micropore volume. The key parameter is the Al M/S ratio, which is the ratio between the number of Al atoms in the side pockets (T_3 sites) and the main channels (T_1, T_2, and T_4 sites). Our model predictions match the computed values very well.

References

(1) Lee, G. J.; Graces, J. M.; Meima, G. R.; Van der Aalst, M. J. M. *U.S. Patent* 5,198,595, 1991.

(2) Tromp, M.; van Bokhoven, J. A.; Oostenbrink, M. T. G.; Bitter, J. H.; de Jong, K. P.; Koningsberger, D. C. *J. Catal.* **2000**, *190*, 209–214.

(3) Leia, G. D.; Carvilla, B. T.; Sachtler, W. M. H. *Appl. Catal. A Gen.* **1996**, *142*, 347–359.

(4) Haag, W. O. In *Zeolites and Related Microporous Materials: State of the Art 1994, Studies in Surface Science and Catalysis*; Weitkamp, J., Karge, H. G., Pfeifer, H., Hölderich, W., Eds.; Elsevier: Amsterdam, The Netherlands, 1994; Vol. 84, pp. 1375−1394.

(5) Schlenker, J. L.; Pluth, J. J.; Smit, J. V. *Mater. Res. Bull.* **1979**, *14*, 751.

(6) Choudhary, V. R.; Mayadevi, S.; Singh, A. P. *J. Chem. Soc. Faraday Trans.* **1995**, *91*, 2935–2944.

(7) Webster, C. E.; Cottone, A.; Drago, R. S. *J. Am. Chem. Soc.* **1999**, *121*, 12127–12139.

(8) Beerdsen, E.; Dubbeldam, D.; Smit, B.; Vlugt, T. J. H.; Calero, S. *J. Phys. Chem. B* **2003**, *107*, 12088–12096.

(9) Marie, O.; Massiani, P.; Thibault-Starzyk, F. *J. Phys. Chem. B* **2004**, *108*, 5073–5081.

(10) Alberti, A. *Zeolites* **1997**, *19*, 411–415.

(11) Niwa, M.; Suzuki, K.; Katada, N.; Kanougi, T.; Atoguchi, T. *J. Phys. Chem. B* **2005**, *109*, 18749–18757.

(12) Bevilacqua, M.; Busca, G. *Catal. Commun.* **2002**, *3*, 497–502.

(13) García-Pérez, E.; Dubbeldam, D.; Liu, B.; Smit, B.; Calero, S. *Angew. Chem., Int. Ed.* **2007**, *46*, 276–278.

(14) Liu, B.; García-Pérez, E.; Dubbeldam, D.; Smit, B.; Calero, S. *J. Phys. Chem. C* **2007**, *11*, 10419–10426.

(15) Leroy, F.; Jobic, H. *Chem. Phys. Lett.* **2005**, *406*, 375–380.

(16) Fan, J. F.; van de Graaf, B.; Xiao, H. M.; Njo, S. L. *J. Mol. Struct.* **1999**, *492*, 133–142.

(17) Fan, J. F.; Wang, Q. X.; Gong, X. D.; Xiao, H. M. *J. Mol. Struct.* **2003**, *638*, 129–134.

(18) Masuda, T.; Fujikata, Y.; Nishida, T.; Hashimoto, K. *Microporous Mesoporous Mater.* **1998**, *23*, 157–167.

(19) Demuth, T.; Hafner, J.; Benco, L.; Toulhoat, H. *J. Phys. Chem. B* **2000**, *104*, 4593–4607.

(20) Macedonia, M. D.; Moore, D. D.; Maginn, E. J.; Olken, M. M. *Langmuir* **2000**, *16*, 3823–3834.

(21) Baerlocher, C.; McCusker, L. B.; Olson, D. H. *Atlas of Zeolite Framework Types*, 6th ed.; Elsevier: Amsterdam, The Netherlands, 2007.

(22) Siepmann, J. I.; Frenkel, D. *Mol. Phys.* **1992**, *75*, 59–70.

(23) Siepmann, J. I. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Ed.; Escom Science Publisher; Leiden, The Netherlands, 1993; pp. 249−264.

(24) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*, 2nd ed. Academic Press: San Diego, CA, 2002.

(25) Smit, B. *Mol. Phys.* **1995**, *85*, 153–172.

(26) Vlugt, T. J. H.; Krishna, R.; Smit, B. *J. Phys. Chem. B* **1999**, *103*, 1102–1118.

(27) Dubbeldam, D.; Calero, S.; Vlugt, T. J. H.; Krishna, R.; Maesen, T. L. M.; Smit, B. *J. Phys. Chem. B* **2004**, *108*, 12301–12313.

(28) Calero, S.; Dubbeldam, D.; Krishna, R.; Smit, B.; Vlugt, T. J. H.; Denayer, J. F. M.; Martens, J. A.; Maesen, T. L. M. *J. Am. Chem. Soc.* **2004**, *126*, 11377–11386.

(29) Vlugt, T. J. H.; Schenk, M. *J. Phys. Chem. B* **2002**, *106*, 12757–12763.

(30) Rapaport, D. C. *The Art of Molecular Dynamics Simulation*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 2004.

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

(32) Krishna, R. *Chem. Phys. Lett.* **2000**, *326*, 477–484.

(33) Dubbeldam, D.; Snurr, R. Q. *Mol. Sim.* **2007**, *33*, 305–325.

(34) Smit, B.; Maesen, T. L. M. *Chem. Rev.* **2008**, *108*, 4125–4184.

(35) Dubbeldam, D.; Ford, D. C.; Ellis, D. E.; Snurr, R. Q. *Mol. Simul.* **2009**; doi: 10.1080/08927020902818039.

(36) Rouquerol, J.; Avnir, D.; Fairbridge, C. W.; Everett, D. H.; Haynes, J. H.; Pernicone, N.; Ramsay, J. D. F.; Sing, K. S. W. *Pure Appl. Chem.* **1994**, *66*, 1739–1758.

(37) Ban, S.; Vlugt, T. J. H. *Mol. Simul.* **2009**; doi: 10.1080/08927020802660614.

Alkanes in Na-MOR

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2865**

(38) Smit, B.; den Ouden, C. J. J. *J. Phys. Chem.* **1988**, *92*, 7169–7171.

(39) Coughlan, B.; Caroll, W. M.; Mcann, W. A. *J. Chem. Soc. Faraday Trans.* **1977**, *73*, 1612.

(40) Mortier, W. J. *Compilation of Extraframework Sites in Zeolites*; Butterworth: Guildford, U.K., 1982.

(41) Devautour, S.; Adboulaye, A.; Giuntini, J. C.; Henn, F. *J. Phys. Chem. B* **2001**, *105*, 9298–9301.

(42) Nesterenko, N. S.; Thibault-Starzyk, F.; Montouillout, V.; Yuschenko, V. V.; Fernandez, C.; Gilson, J. P.; Fajula, F.; Ivanova, I. I. *Microporous Mesoporous Mater.* **2004**, *71*, 157–166.

(43) Katada, N.; Takeguchi, T.; Suzuki, T.; Fukushima, T.; Inagaki, K.; Tokunaga, S.; Shimada, H.; Sato, K.; Oumi, Y.; Sano, T.; Segawa, K.; Nakai, K.; Shoji, H.; Wu, P.; Tatsumi, T.; Komatsu, T.; Masuda, T.; Domen, K.; Yoda, E.; Kondo, J. N.; Okuhara, T.; Kageyama, Y.; Niwa, M.; Ogura, M.; Matsukata, M.; Kikuchi, E.; Okazaki, N.; Takahashi, M.; Tada, A.; Tawada, S.; Kubota, Y.; Sugi, Y.; Higashio, Y.; Kamada, M.; Kioka, Y.; Yamamoto, K.; Shouji, T.; Arima, Y.; Okamoto, Y.; Matsumoto, H. *Appl. Catal. A Gen.* **2005**, *283*, 63–74.

(44) Pamba, M.; Maurin, G.; Devautour, S.; Vanderschueren, J.; Giuntini, J. C.; Renzo, F. D.; Hamidi, F. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2027–2031.

(45) Tyburce, B.; Kappenstein, C.; Cartraud, P.; Garnier, E. *J. Chem. Soc. Faraday Trans.* **1991**, *87*, 2849–2853.

JCTC Journal of Chemical Theory and Computation

Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional

Carlos Campañá, Bastien Mussard, and Tom K. Woo*

Centre for Catalysis Research and Innovation and Department of Chemistry, University of Ottawa, Ottawa, Canada, K1N 6N5

Abstract: A method to generate electrostatic potential (ESP) derived atomic charges in crystalline solids from periodic quantum mechanical calculations, termed the REPEAT method, is presented. Conventional ESP fitting procedures developed for molecular systems, in general, will not work for crystalline systems because the electrostatic potential in periodic systems is ill-defined up to a constant offset at each spatial position. In this work the problem is circumvented by introducing a new error functional which acts on the relative differences of the potential and not on its absolute values, as it is currently done with molecular ESP charge derivation methods. We formally demonstrate that the new functional reduces to the conventional error functional used in molecular ESP approaches when the simulation box of the periodic calculation becomes infinitely large. Several tests are presented to validate the new technique. For the periodic calculation of isolated molecules, the REPEAT charges are found to be in good agreement with those determined with established molecular ESP charge derivation methods. For siliceous sodalite, it is demonstrated that conventional molecular ESP approaches generate 'unphysical' charges, whereas the REPEAT method produces charges that are both chemically intuitive and consistent between different periodic electronic structure packages. The new approach is employed to generate partial atomic charges of various microporous materials and compared to both experimentally derived and molecular fragment ESP charges. This method can be used to generate partial atomic charges to be used in simulations of microporous and nanoporous materials, such as zeolites and metal organic framework materials.

I. Introduction

Microporous materials are a fascinating class of condensed matter systems with a wide range of important applications. They are commonly employed as catalysts in the petrochemical industry, as adsorbents to trap impurities, as physical reservoirs for gas storage, as molecular sieves for size and shape selective separations, and as mediums for ion exchange processes.[1] Perhaps the most well-known microporous materials are zeolites, which are crystalline solids with well-defined structures and pores that range in size from $\sim 1-20$ Å.[1] Recently, a new class of microporous materials known as metal organic frameworks (MOFs)[2] have emerged that are composed of metal ions and bridging organic ligands. Compared to zeolites, MOFs have a much wider variety of

structural and chemical motifs that promise a higher degree of tunability. As such, MOFs have attracted significant attention for their potential applications in hydrogen storage and CO_2 sequestration.[3−9]

Atomistic modeling of microporous materials, particulary zeolites, have contributed enormously to increase our understanding on their properties and functionality.[10,11] Most studies of the full periodic systems have employed empirical interatomic potentials (force fields), although periodic density functional theory (DFT) studies have recently emerged.[11] Molecular dynamics and Monte Carlo simulations using empirical potentials have been successfully applied to investigate the dynamics of guest molecules as well as their adsorption characteristics within microporous materials.[10−13] In some cases, atomistic simulations have led to the design of improved compounds with enhanced properties and

---

* Corresponding author. E-mail: twoo@uottawa.ca.

Electrostatic Potential Derived Atomic Charges

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2867**

performance.[14,15] Although several generally transferable force fields have been developed for organic molecules, the treatment of electrostatic interactions still remains a challenge when employing empirical potentials. Most force fields use fixed partial atomic charges to treat the electrostatic interactions, although more sophisticated methods have been developed but are not in wide use. Since there are no 'true' atomic charges within polynuclear systems, an assortment of charge derivation methods have been created for various quantitative and descriptive purposes. The so-called electrostatic potential (ESP) derived charges are most commonly used for atomistic simulation of molecular systems.[16−19] In order to compute ESP charges, a quantum chemical calculation is performed on a molecule, and partial atomic charges are fit to reproduce the quantum chemical electrostatic potential on a fine grid surrounding the molecule. The grid is chosen to lie outside of the van der Waals radius (or similar) of each atom in the molecule. The charges are 'designed' to reproduce the electrostatic potential in the region where it is most important when modeling intermolecular interactions with two-body, additive Coulomb potentials. In situations where there are deeply embedded (buried) atoms in the molecule, the ESP charges for these buried atoms can fluctuate widely, sometimes resulting in nonchemically intuitive values. To deal with this problem, Bayly and co-workers introduced the RESP method,[18] where a penalty function is introduced to the fitting procedure to inhibit the 'unphysical' charges from arising. ESP charges have been shown to provide superior results compared to those using other schemes, such as charges derived from a Mulliken population analysis, and therefore, ESP charges are the norm for molecular simulation when employing point charge electrostatic models.[16,18]

For the atomistic simulation of periodic systems, ESP charges are not as universally applied as they are for molecular systems. In highly packed solids, where there is little volume outside of the van der Waals radii of the atoms to define valid fitting points, ESP charges may not be appropriate. However, even for the simulation of microporous materials where there are large pore volumes, Mulliken charges derived from periodic DFT calculations are still in wide usage despite the strong basis set dependence of the scheme.[13,20] If ESP charges are used for periodic simulations, they are derived by extracting a fragment of the periodic lattice and performing a molecular ESP calculation.[12,21−23] The assumption with this procedure is that the electrostatic potential surrounding the isolated fragment is going to be similar to that of the periodic structure. Considering the care taken to treat long-range electrostatic interactions when using periodic boundary conditions in molecular simulations, this assumption may not be generally valid. Additionally, special considerations also have to be made to 'cap' the electronic system to satisfy unfilled valences resulting from the extraction or account for the net charge of the retrieved fragment. Finally, the properties of a periodic wave function including the electrostatic potential are averages over the Brillouin zone. Molecular fragment calculations cannot capture the effect of proper K-point sampling, which is important for some periodic systems.
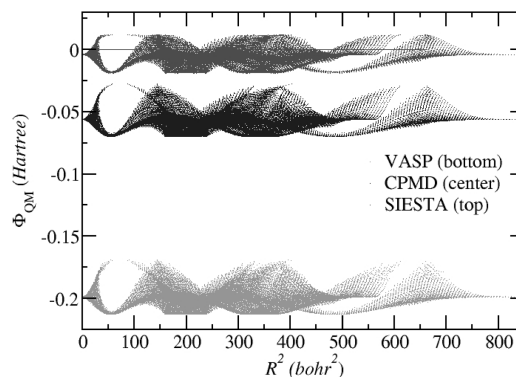

Figure 1. Electrostatic potential plotted as a function of the square of the distance from the cell origin for sodalite generated from three periodic DFT packages: VASP (bottom), CPMD (center), and SIESTA (top). The tabulation is performed on a cubic grid, and only points located outside the van der Waals radii centered on each atom are shown.

Although periodic 'first principles' (i.e., DFT) calculations of materials with large unit cells containing hundreds of atoms are now routinely performed, to the best of our knowledge, a method to derive charges from the electrostatic potential of periodic first principles calculations has not been reported.[24,25] One of the reasons for the dearth of development of periodic ESP fitting techniques, as compared to molecular systems, may be due to the fact that the absolute energy of an atom is not an intrinsic bulk property, and therefore, the average electrostatic potential is ill-defined.[26] In other words, the reference state of the electrostatic potential in a periodic system is arbitrary. This is numerically demonstrated in Figure 1, where the electrostatic potential for siliceous sodalite (a zeolite) is plotted as a function of the square distance to the cell origin on a mesh resulting from three different periodic DFT packages − VASP, CPMD, and SIESTA.[27] Qualitatively, the general structure of the electrostatic potentials show the same features, except that the manifold derived from each periodic DFT package is shifted by a different but constant offset amount. Since the average of the electrostatic potential in an infinite system is ill-defined, there is no 'correct' offset value, and the straightforward application of the ESP charge fitting procedures developed for molecular systems will result in charges that are dependent on this arbitrary offset value.

In this work, we introduce a modified ESP error functional that can be used in periodic systems as well as in molecular systems. Our functional allows for a robust computation of the ESP charges while overcoming the fundamental problem of the electrostatic potential having an ill-defined reference state. We also introduce a new RESP-like penalty function to prevent large fluctuations in the fitted charges of buried atoms that is based on the expansion of the energy of an atom as a function of charge. In the tradition of ESP charge methods, we also suggest an acronym for the charges derived from the method. We refer to these charges as REPEAT (Repeating Electrostatic Potential Extracted ATomic) charges. The remainder of the paper is as follows: The formalism of the method is introduced in detail with the appropriate equations in Section II. Section III describes the choice of parameters used in our implementation of the REPEAT

method as well as details on the periodic DFT calculations that were performed to generate the electrostatic potentials. Section IV contains the tests and the applications used to validate the REPEAT method, and Section V concludes the paper.

II. Methodology

For molecular systems, the electrostatic potential at a point $\vec{r}$ that results from a set of N_q point charges, $\{q_j\}$, is given in eq 1 (in atomic units), where $\vec{r}_j$ represents the position of each of the point charges:

$$\phi_q(\vec{r}) = \sum_j^{N_q} \frac{q_j}{|\vec{r} - \vec{r}_j|} \tag{1}$$

In generating ESP charges for molecular systems, a quantum chemical calculation is first performed to generate a reference electrostatic potential, $\phi_{QM}(\vec{r})$. The set of point charges $\{q_j\}$ is then adjusted to minimize the differences between the 'quantum mechanical' electrostatic potential, $\phi_{QM}(\vec{r})$, and the electrostatic potential due to the atomic charges defined by eq 1, for grid points surrounding the atoms in the molecule. This is typically performed in a least-squares manner where the following function is minimized:

$$F(\{q_j\}) = \sum_{grid} (\phi_{QM}(\vec{r}_{grid}) - \phi_q(\vec{r}_{grid}))^2 + \lambda(\sum_j q_j - q_{tot}) \tag{2}$$

The second term in eq 2 represents a Lagrange multiplier that ensures that the sum of atomic charges equals the total charge of the system, q_{tot}. The grid points used for the fitting procedure are chosen to lie outside of the van der Waals (VDW) radii (or similar) of the atoms in the molecule. The actual algorithm for choosing the grid points varies between different methods, and several ESP charge fitting techniques are in wide usage, such as the CHELPG[17] and the RESP[18] approaches.

When generating ESP derived atomic charges for periodic systems, there are additional complications that need to be addressed as compared to molecular ESP charges. First, the electrostatic potential due to the point charges has to account for the infinite periodic images of the atomic charges. The Ewald summation technique is a well-known mathematical procedure to deal with the long-range electrostatic interactions in periodic systems. Within the Ewald procedure, the electrostatic potential is factored in two terms, real and reciprocal space series. The real space series accounts for the effects of the short-range interactions. Conversely, the reciprocal space term includes the long-range effects arising from the presence of the infinite images. For a formal derivation of the mathematical expressions representing each one of the series, we refer the reader to the original and related works.[28,29] Summarizing, the exact expression for the potential according to the Ewald formulation is given in eq 3 where the erfc stands for the complementary error function.

$$\phi_q(\vec{r}) = \sum_{j,T} q_j \frac{erfc(\alpha|\vec{r} - \vec{r}_{j,T}|)}{|\vec{r} - \vec{r}_{j,T}|} + \frac{4\pi}{V} \sum_{j,\vec{k}} q_j e^{i\vec{k}(\vec{r}-\vec{r}_j)} \frac{e^{-\frac{k^2}{4\alpha^2}}}{k^2} \tag{3}$$

The vectors $\vec{T} = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3$, appearing in the first double summation, map the real space lattice positions. Conversely, the vectors $\vec{k} = m_1\vec{b}_1 + m_2\vec{b}_2 + m_3\vec{b}_3$, included within the second double summation, map the reciprocal space lattice. The variable α represents the width of the screening Gaussian charge distribution added to each atomic center. Equation 3 can be considered the periodic analogue of eq 1. With infinite terms, eq 3 is mathematically exact and independent of the value chosen for the parameter α. However, the value of α as well as the number of terms in the real and reciprocal space series determine the speed of convergence and accuracy of the numerical computation of the electrostatic potential. Techniques to select optimum values for these parameters have been discussed elsewhere.[30] If the charges are to be used in a molecular dynamics simulation, then it can be argued that these Ewald summation parameters used to derive the ESP charges should ideally be the same as those used in the simulations.

The second complication when deriving periodic ESP charges relates to the ill-defined nature of the reference state of the electrostatic potential, as depicted in Figure 1. Even if eq 3 is used for ϕ_q, minimization of the conventional error function (eq 2) developed for molecular systems will not be of any value when attempting to compute charges in periodic systems. For example, it is clear that the different electrostatic potentials shown in Figure 1 will yield drastically different fitted charges. (There are reports of conventional ESP charge fitting procedures yielding unsatisfactory results for periodic systems.)[31,32] To overcome this fundamental problem, we introduce a modified error functional to perform the least-squares fit. Our modified error functional can be expressed as the sum of two independent terms:

$$F(\{q_j, \delta_\phi\}) = \sum_{grid} (\phi_{QM}(\vec{r}_{grid}) - (\phi_q(\vec{r}_{grid}) + \delta_\phi))^2 + \\ \lambda(\sum_j q_j - q_{tot}) \tag{4}$$

where the new parameter δ_ϕ is given by

$$\delta_\phi = \frac{1}{N_{grid}} \sum_{grid} (\phi_{QM}(\vec{r}_{grid}) - \phi_q(\vec{r}_{grid})) \tag{5}$$

As in molecular approaches, the second term in eq 4 represents a Lagrange multiplier that ensures that the total charge of the system remains equal to q_{tot}, which is zero in the case of periodic systems. Examination of the definition of δ_ϕ in eq 5 reveals that the ESP fit is performed such that only the relative differences with respect to the average value of the potential on the grid points are considered. This automatically guarantees that the quality of the fit is not affected by a constant offset in the estimation of the potential values. It is important to note that the expression for δ_ϕ can be obtained in a simple and straightforward manner by assuming δ_ϕ in eq 4 is another independent variable in the error functional $F(\{q_j, \delta_\phi\})$. The definition of δ_ϕ in eq 5 can

Electrostatic Potential Derived Atomic Charges

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2869**

then be derived by performing the minimization of F not only with respect to the charges but also with respect to δ_ϕ.

Explicit consideration of δ_ϕ within the error functional form not only removes the artifact of an ill-defined reference state but implicitly introduces corrections to the dipole moment and the other higher-order multipoles, if our technique is to be employed to fit charges in molecular compounds. More specifically, it is easy to show that for a neutral molecular system if one expands δ_ϕ in the continuum representation up to the first-order (dipole) corrections one obtains

$$\delta_\phi = \frac{1}{N_{\text{grid}}} \sum_{\text{grid}} (\phi_{\text{QM}}(\vec{r}_{\text{grid}}) - \phi_q(\vec{r}_{\text{grid}})) \approx \Delta\phi_{q-\text{QM}} + \frac{1}{V} \int \frac{(\vec{p}_{\text{QM}} - \vec{p}_q)\hat{r}}{|\vec{r}|^2} d^3r \quad (6)$$

where $\vec{p}$ is the electric dipole moment and $\Delta\phi_{q-\text{QM}}$ is just the difference between the zero levels of the quantum and the classical charge distribution, which is expected to be zero in molecular systems. The second term reveals that errors in the dipole moment can be lowered using our fitting technique. The possibility of including implicit corrections to multipole moments allows for capturing potential polarization effects in the cluster calculations of charges within highly polar molecular systems.

In systems in which there are buried atoms, such that there are few valid fitting grid points close to these atoms, the ESP charge fitting approach can yield charges that are considered 'unphysical' or chemically nonintuitive. To avoid wide fluctuations in the fitted charges for buried atoms, penalty functions are added to the error functional, such as those used in the RESP method.[18] In order to generalize our methodology to systems in which buried atoms could be present, we propose adding a second set of Lagrange multipliers, w_j, such that the REPEAT error functional becomes:

$$F(\{q_j, \delta_\phi\}) = \sum_{\text{grid}} (\phi_{\text{QM}}(\vec{r}_{\text{grid}}) - (\phi_q(\vec{r}_{\text{grid}}) + \delta_\phi))^2 + \lambda(\sum_j q_j - q_{\text{tot}}) + \sum_j w_j\left(E_j^0 + \chi_j q_j + \frac{1}{2}J_j^{00}q_j^2\right) \quad (7)$$

As opposed to the original RESP method and derivative works,[18] we have made no ad hoc assumptions regarding the shape of the penalty term. We use a physically motivated penalty multiplier that is an expansion of the energy of an atom (idem to the potential in atomic units) as a function of the atomic charges up to the second order.[33,34] The parameters χ_j and J_j^{00} are the electronegativity and the self-Coulomb interaction, respectively, of the individual chemical elements. The nature of the self-Coulomb interaction term J_j^{00} and how to determine it is discussed in more detail by Rappe and Goddard.[34] The free parameter included in the model is w_j, which can be considered a weighting factor used to adjust the strength of the restraints. A single weighting factor w could be used for all atoms, or as expressed in eq 7, the weighting factors can be individually adjusted to selectively turn on penalty corrections only for atoms identified as buried prior to performing the charge calculations. Aside from being

a physically motivated restraining method, the method also has the advantage that the derivative of the error function with respect to the charges is linear (vide infra), and therefore, there is no need to solve the equations iteratively as is required with the hyperbolic constraints of RESP.

In the following, we provide the full derivation of the linear system of equations to be solved in order to calculate atomic charges from the spatial tabulation of the periodic electrostatic potential within the REPEAT approach. Minimizing the error functional, F, defined in eq 7 with respect to the independent variables $\{q_j\}$ and δ_ϕ yields the set of equations:

$$\frac{\partial F}{\partial q_{j=1,\ldots,N_q}} = \begin{array}{l} -2\sum_{\text{grid}} (\phi_{\text{QM}}(\vec{r}_{\text{grid}}) - \delta_\phi - \phi_q(\vec{r}_{\text{grid}})) \times \\ \left(\frac{\partial\phi_q(\vec{r}_{\text{grid}})}{\partial q_j} + \frac{\partial\delta_\phi}{\partial q_j}\right) + \lambda + w_j(\chi_j + J_j^{00}q_j) = 0, \end{array}$$

$$\frac{\partial F}{\partial \delta_\phi} = -2\sum_{\text{grid}} (\phi_{\text{QM}}(\vec{r}_{\text{grid}}) - \delta_\phi - \phi_q(\vec{r}_{\text{grid}})) = 0 \quad (8)$$

The last equation above defines δ_ϕ as presented in eq 5. Using the definition of the derivative of the potential with respect to the charges:

$$\frac{\partial\phi_q(\vec{r}_{\text{grid}})}{\partial q_j} = \sum_{\vec{T}} \frac{\text{erfc}(\alpha|\vec{r}_{\text{grid}} - \vec{r}_{j,\vec{T}}|)}{|\vec{r}_{\text{grid}} - \vec{r}_{j,\vec{T}}|} + \frac{4\pi}{V}\sum_{\vec{k}} e^{i\vec{k}(\vec{r}_{\text{grid}} - \vec{r}_j)}\frac{e^{-\frac{k^2}{4\alpha^2}}}{k^2} = \phi_j'(\vec{r}_{\text{grid}}) \quad (9)$$

combined with the first group of eq 8 and after some lengthy algebra one arrives at a matrix problem with $N_q + 1$ unknowns of the type:

$$\mathbf{A}\tilde{\mathbf{Q}} = \mathbf{B} \quad (10)$$

with matrix coefficients $\mathbf{A}_{jm}$ defined as

$$\mathbf{A}_{jm,(j,m)<N_q+1} = \sum_{\text{grid}}\left\{\left(\phi_j'(\vec{r}_{\text{grid}}) - \frac{1}{N_{\text{grid}}}\sum_{\text{grid*}} \phi_j'(\vec{r}_{\text{grid*}})\right) \times \right. \\ \left. (\phi_m'(\vec{r}_{\text{grid}}) - \frac{1}{N_{\text{grid}}}\sum_{\text{grid*}} \phi_m'(\vec{r}_{\text{grid*}}))\right\} + \frac{w_j}{2}J_j^{00}\delta_{jm} \quad (11)$$

and

$$\mathbf{A}_{jm,j<N_q+1,m=N_q+1} = \mathbf{A}_{jm,j=N_q+1,m<N_q+1} = \mathbf{A}_{jm,j=m=N_q+1} = 1 - \delta_{jm} \quad (12)$$

The vector $\mathbf{B}$ on the right-hand side of eq 10 has elements

$$\mathbf{B}_{m,m<N_q+1} = \sum_{\text{grid}}\left\{(\phi_{\text{QM}}(\vec{r}_{\text{grid}}) - \frac{1}{N_{\text{grid}}}\sum_{\text{grid*}} \phi_{\text{QM}}(\vec{r}_{\text{grid*}}))(\phi_m'(\vec{r}_{\text{grid}}) - \frac{1}{N_{\text{grid}}}\sum_{\text{grid*}} \phi_m'(\vec{r}_{\text{grid*}}))\right\} - w_m\frac{\chi_m}{2} \quad (13)$$

and

$$\mathbf{B}_{m,m=N_q+1} = q_{\text{tot}} \quad (14)$$

The elements of the unknown vector $\tilde{\mathbf{Q}}$ are given by:

$$\tilde{\mathbf{Q}}_{m,m<N_q+1} = q_m \qquad (15)$$

and

$$\tilde{\mathbf{Q}}_{m,m=N_q+1} = \frac{\lambda}{2} \qquad (16)$$

In the case where the crystallographic symmetry of the system demands the equivalency of some of the atomic sites and therefore charges, a minor modification needs to be added to the above formulation. The columns and rows corresponding to identical sites in the **A** matrix (rows in the case of the right-hand side vector **B**) must be combined into a single unit by adding their elements.

III. Computational Details

Due to the numerical nature of this work, in this section we provide a discussion of the parameters chosen in the different tests that were performed in order to validate our methodology. We also provide the details of the DFT calculations used to generate the electrostatic potentials of the different systems considered as well as the details of other quantum chemical calculations used for comparison purposes. Some information on our implementation and coding of the REPEAT method are also included.

For the Ewald summations, the real space cutoff has been fixed to be $R_{cut} = 9$ Å, and the reciprocal space cutoff order has been fixed to be $k_{max} = 7$. The magnitude for the width of the Gaussian screening charge distribution was set as $\alpha = \sqrt{\pi}/R_{cut}$. These are typical values of the Ewald summation parameters recommended in classical molecular dynamics packages. We note that, when calculating the contributions of the real and reciprocal space series, the speed of convergence is not an issue because such series only have to be computed one time during the fitting process.

Only those grid points located outside the VDW spheres centered on each atomic site have been considered. Atomic regions due to the periodic images of the atoms are also excluded. The VDW radii used for this work were taken from the universal force field (UFF).[35] The grid points used for the fitting procedure coincided with the regular numerical FFT grid used in the periodic DFT calculation unless otherwise specified. The majority of the REPEAT charges reported in this work are unrestrained. In other words, the last term of the REPEAT error functional of eq 7 is set to zero. For results in which restraints were utilized, the electronegativity, χ, and self-Coulomb interaction parameters, J^{00}, of ref 34 are used.

Periodic DFT calculations were performed with several packages, VASP,[36] CPMD,[37] Gaussian 03,[38] and SIESTA.[39] All DFT calculations used the Perdew−Burke−Ernzerhof (PBE)[40] exchange and correlation functional. For the VASP calculations, the PAW method of Blöchl[41,42] was utilized with a plane wave cutoff of 600 eV. For the CPMD calculations the Goedecker−Teter−Hutter pseudopotentials[43] were used with a plane wave cutoff of 80 Ryd (∼1088 eV). The Brillouin zone integrations were performed using a 4 × 4 × 4 Monkhorst−Pack grid for the calculations of

sodalite, natrolite, and zinc silicate, while a 14 × 14 × 14 sampling of the Brillouin zone was used for SnO$_2$, ZnO, and CdTe. For calculations of the isolated molecules and IRMOF-1, only the Γ-point was evaluated. True molecular calculations and CHELPG charges were performed with the Gaussian 03 package. RESP charges were generated using the AMBER 8 package.[44] To determine atomic charges using the Gasteiger's electronegativity equalization method[45] and Rappé's QEq method,[34] the Materials Studio modeling package version 4.1 was used. Except for the calculation of the isolated molecules, experimental geometries were used in evaluating the electrostatic potential.

We have implemented the REPEAT method as a FORTRAN code that is freely available upon request from the authors. The program reads a tabulation of the electrostatic potential in the Gaussian cube file format and can handle any unit cell shape.

To compare the electrostatic potential resulting from the fitted charges to that generated from the DFT calculation, we use the relative root-mean-square (RRMS) error defined in eq 17, where the sum runs over valid grid points or those lying outside the van der Waals radii of the atoms.

$$RRMS = \sqrt{\frac{\sum\limits_{grid} (\phi_{QM}(\vec{r}_{grid}) - \phi_q(\vec{r}_{grid}))^2}{\sum\limits_{grid} \phi_{QM}(\vec{r}_{grid})^2}} \qquad (17)$$

IV. Results and Discussion

1. Charges Derived From Different DFT Packages. As a first validation of the REPEAT method, we first compare the charges derived from the electrostatic potential of sodalite depicted in Figure 1 using the conventional error functional of eq 2 and the charges derived using the REPEAT method with its modified error functional given in eq 4 (without the restraints).

Sodalite is a zeolite whose silicate framework is composed of tetrahedrally coordinated Si atoms linked through two coordinated oxygen atoms. This material has a cubic structure with a lattice parameter $a = 8.965$ Å and belongs to the $P\bar{4}3n$ space group. There are 36 atoms (12 Si, 24 O) in the unit cell, with only one symmetry unique Si and O atom. Three periodic DFT packages VASP, CPMD, and SIESTA were used to compute the electrostatic potential within the unit cell on a regular grid, using the PBE exchange correlation functional. The tabulation of the potential as a function of the square distance to the unit cell origin is shown in Figure 1, as already mentioned. The grid point density from each of the DFT calculations varied, but the total number of grid points available to perform the fitting (those outside of the VDW radii) was on the order of 3×10^5 for all cases.

Given in Table 1 are the ESP derived charges using the conventional error function of eq 2. The charges fluctuate widely depending on the DFT package used, and in some cases, the charges lack chemical meaning. It is notable that the RRMS errors also vary widely reaching up to 43% in the case of SIESTA. In contrast, the REPEAT charges also given in Table 1 are rather insensitive to the DFT code,

Electrostatic Potential Derived Atomic Charges

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2871**

Table 1. ESP Charges of Sodalite Derived from the Electrostatic Potential Generated from Various Periodic DFT Packages Using the Standard Error Function and with The REPEAT method

atom	conventional ESP charges			REPEAT ESP charges		
	CPMD	VASP	SIESTA	CPMD	VASP	SIESTA
Si	+3.075	+10.522	+0.503	+1.355	+1.389	+1.151
O	−1.537	−5.261	−0.251	−0.677	−0.694	−0.575
RRMS error	0.193	0.289	0.426	0.066	0.066	0.064

Table 2. Comparison between REPEAT, CHELPG, and RESP ESP Charges for Isolated Ammonia, Water, And Alanine Molecules

molecule	atom	CHELPG[a]	RESP[a]	REPEAT[b]	REPEAT (restrained)[c]
NH_3	N	−0.956	−0.752	−0.814	−0.813
	H	+0.319	+0.251	+0.271	+0.271
H_2O	O	−0.706	−0.704	−0.661	−0.661
	H	+0.353	+0.352	+0.330	+0.330
$CH_3NH_2CHCOOH$	C	−0.289	−0.321	−0.329	−0.318
	H	+0.075	+0.098	+0.091	+0.088
	H	+0.091	+0.098	+0.109	+0.107
	H	+0.083	+0.098	+0.098	+0.095
	N	−0.848	−0.795	−0.814	−0.805
	H	+0.334	+0.329	+0.322	+0.319
	H	+0.342	+0.329	+0.329	+0.327
	C	+0.328	+0.233	+0.264	+0.261
	H	+0.089	+0.022	+0.007	+0.006
	C	+0.530	+0.508	+0.612	+0.610
	O	−0.482	−0.464	−0.523	−0.523
	O	−0.462	−0.446	−0.481	−0.480
	H	+0.318	+0.311	+0.316	+0.314

[a] ESP charges derived from a molecular PBE/6-31G(d,p) calculation. [b] ESP charges derived from a VASP calculation of the molecules using cubic supercell of length 20 Å and the PBE functional. [c] Calculated with the restraining function introduced in eq 7.

although, small variations are to be expected since the three codes are using different types of basis sets and pseudopotentials. Table 1 not only shows that there are only small variations in the charges from one code to another, but it also shows that the root-mean-square errors remain quite similar and significantly smaller than those computed with the conventional functional.

To complement the numerical example above, it is also of value to consider a straightforward mind experiment to test the robustness of the REPEAT approach. In this experiment, it is assumed the existence of an ideal neutral periodic point charge distribution for which all the charges are known a priori. However, the tabulation of its corresponding electrostatic potential has been modified such that a small systematic constant error, δ_e, has been added to each one of its local values. If both the conventional as well as our REPEAT functional were to be applied on such a redefined potential, one can intuitively guess that the only way of recuperating the original values of the point charges is by using the REPEAT method. Moreover, the REPEAT method should provide correct charges regardless of the magnitude of the error level δ_e. Conversely, applying the conventional definition of the error functional makes the predictions for the charges level dependent and does not render the right values unless the error in the level is identically zero.

The REPEAT charges for sodalite can be compared to those derived for sodalite and related silicates using other methods where there is wide variation depending on the approach used. Nicholas and co-workers[46,47] developed charges of +1.1 e for Si and −0.55 e for O to model sodalite. These charges were derived from fragment calculations and adjusted to intrinsically account for cations in the sodalite framework when Si is replaced by Al atoms in the framework. For silicates, Si charges ranging from +0.4 to +1.91 e have been utilized.[47] Mulliken charges using the traditional atom centered basis functions gave a very wide range of charges for sodalite. Specifically, charges on Si were calculated to be +1.693, +1.723, and +0.965 e for the 3-21G, 6-31G, and 6-31G(d,p) basis sets, respectively. Using Gasteiger's electronegativity equalization method[45] gives charges of +0.624 e on Si and −0.312 e for O, while the QEq method of Rappé and Goddard[34] provides charges of +1.00 e on Si and −0.50 e on O.

2. Molecular Calculations. The previous results given in Table 1 reveal that the REPEAT algorithm minimizes the dependency of the charges with respect to the periodic DFT code. Although they show that similar charges are derived regardless of the arbitrary offset in the electrostatic potential,

a natural question arises. Does the REPEAT method simply provide an optimum least-squares fit of the problem or do the charges generated represent more pragmatic charges similar in the spirit of existing ESP charge derivation methods? To examine this question, we study the REPEAT charges obtained from isolated molecule calculations and compare them to established molecular ESP charge derivation methods.

Provided in Table 2 are the CHELPG and RESP charges of ammonia, water, and alanine resulting from a PBE/6-31G(d,p) isolated molecule calculation. Also presented in Table 2 are the REPEAT charges resulting from periodic VASP calculations of the molecules, using the same geometry as the molecular calculations. In the periodic calculations, a simple cubic simulation cell of $L = 20$ Å was used to minimize the interaction with the periodic images. Table 2 reveals that the CHELPG, RESP, and REPEAT charges are all in good agreement with one another. One should note that the CHELPG and RESP charges were derived from the same electrostatic potential, whereas the REPEAT charges were derived from a separate DFT calculation using a different basis set type and pseudopotential.[48] It is further notable that the total number of grid points (not shown in the table) used for the REPEAT charges was almost 2 orders of magnitude greater than those of the CHELPG and RESP methods.

Also provided in Table 2 are the REPEAT charges in which the quadratic restraints first introduced in eq 7 are applied. We have used a weighting factor defined as $w_j = 2a/J_j^{00}$ for each atom, where a is the weighting factor employed within the parabolic RESP approach. This is equivalent to using the weighting factor of 0.005 recommended by Baley et al. for the quadratic form of the RESP restraint.[18] The changes in the REPEAT charges resulting from the application of the restraints can be characterized as negligible. In these examples there are no embedded

atoms, and such negligible changes are desirable because the restraints will always act to increase the RRMS difference between the quantum mechanical electrostatic potential and the point charges. A detailed study of the most appropriate restraining parameters to evaluate REPEAT charges of periodic systems with embedded atoms will be the subject of an upcoming publication. In this paper, we focus on the general applicability of the REPEAT method to derive ESP charges for periodic systems but include the restraining formulation to retain the complete REPEAT formulation (eqs 7−16) in one place. It is worth mentioning that, in general, the charges predicted in molecular calculations are more susceptible to the RESP-type penalties than those obtained for periodic systems.

The results of Table 2 suggest that for isolated molecule calculations the REPEAT method does generate charges that are consistent with other ESP derived charge methods. Indeed, it can be shown that for the periodic calculation of isolated molecules the charges derived using the conventional ESP error functional converge to the REPEAT charges as the size of the simulation cell increases. If we rewrite the definition of the unrestrained REPEAT functional up to first-order corrections using eq 6 we obtain

$$F(\{q_j, \delta_\phi\}) = \sum_{grid} (\phi_{QM}(\vec{r}_{grid}) - \phi_q(\vec{r}_{grid}) - \Delta\phi_{q-QM} - I(\Delta\vec{p}_{q-QM}))^2 \quad (18)$$

where the new term $I(\Delta\vec{p}_{q-QM})$ is the integral over the dipole differences as appearing in the second term of eq 6. In the limit of an infinitely large box, the term $\Delta\phi_{q-QM}$ becomes the 'reference zero' of the quantum electrostatic potential because the 'reference zero' of the classical potential is strictly null. Therefore eq 18 can be rewritten as

$$F(\{q_j, \delta_\phi\}) = \sum_{grid} (\phi_{QM}^*(\vec{r}_{grid}) - \phi_q(\vec{r}_{grid}) - I(\Delta\vec{p}_{q-QM}))^2 \quad (19)$$

with $\phi_{QM}^* = \phi_{QM} - \Delta\phi_{q-QM}$ representing the 'true' quantum potential with respect to its well-defined zero. Thus, in order to prove that the REPEAT functional reduces to the conventional functional in the limit of an infinitely large box, we only need to show that $I(\Delta\vec{p}_{q-QM})$ approaches zero in such limit. To achieve such a goal, we split the valid grid points into two regions of volume V_1 and V_2 as shown schematically in Figure 2. The first region is that lying outside of the VDW radii of all atoms and enclosed by a sphere centered at the charge origin of the molecule and with a radius R_1. We define R_1 as the distance from the charge origin to the furthest grid point lying on a VDW surface of an atom, where we have assumed that the charge origin is located at the center of a cubic simulation box. This first region of volume V_1 is that shaded in Figure 2. Next, the second volume is defined as that lying outside of the first sphere of radius R_1 but lying within a second sphere centered at the charge origin of radius $R_2 = L/2$, where L is the length of the simulation box. Using the above way of separating the grid into regions, the term $I(\Delta\vec{p}_{q-QM})$ can be written as

$$I(\Delta\vec{p}_{q-QM}) = \frac{1}{V} \int \frac{(\vec{p}_{QM} - \vec{p}_q)\hat{r}}{|\vec{r}|^2} d^3r =$$

$$\frac{1}{(V_1 + V_2)} \int_{V_1} \frac{(\Delta\vec{p}_{q-QM})\hat{r}}{|\vec{r}^2|} d^3r + \frac{1}{(V_1 + V_2)} \int_{V_2} \frac{(\Delta\vec{p}_{q-QM})\hat{r}}{|\vec{r}^2|} d^3r \quad (20)$$

The last integral in eq 20 turns out to be identically zero for symmetry reasons. Notice that the integration takes place on a region located between two spherical surfaces. Alternatively, the first integral yields a finite number, which is independent of the size of the simulation box. Nevertheless, having the first integral being multiplied by the inverse total volume allows us to make the contribution of the full $I(\Delta\vec{p}_{q-QM})$ as small as desired by increasing the simulation box size. In an analogous way, one can show that higher moment contributions could be made as small as desired. Thus, we have analytically shown that the REPEAT charges reduce to the standard ESP charges in the limit of an infinitely large simulation box.

To complement the previous analytical proof, we will numerically demonstrate that for molecular calculations the charges derived using the conventional error functional converge to the REPEAT charges as the size of the simulation cell increases. To do this, we have performed a series of water molecule calculations with VASP in which the cubic simulation cell is varied. At large box sizes, the water molecules can be considered noninteracting. However, as the box size, L, is decreased, the water molecule and its periodic images will begin to interact with one another. This situation is shown schematically in Figure 3. Since the dipole moments of the molecule and its images are aligned, the molecules should become more polarized as the simulation cell length is shortened. In other words, we would expect the fitted charges to become larger in magnitude as the length, L, is diminished.

Plotted in Figure 3 are the results of these calculations for both the REPEAT and the ESP charges derived by using the standard ESP error functional. As expected, the REPEAT charges become larger in magnitude as the simulation box size decreases. On the other hand, the ESP charges using the standard error functional become smaller in magnitude, in contradiction to what should physically occur. This is
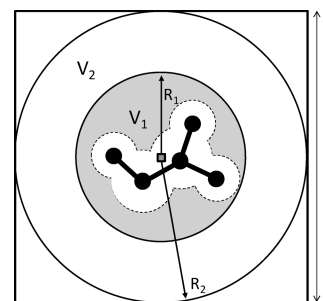

Figure 2. Schematic representation of an isolated molecule within a cubic simulation box of length L, defining the radii R_1 and R_2 of eq 20. The dotted outline represents the van der Waals radii of the atoms within the molecule. The shaded area represents V_1. The charge origin of the molecule is denoted by the square.

Electrostatic Potential Derived Atomic Charges

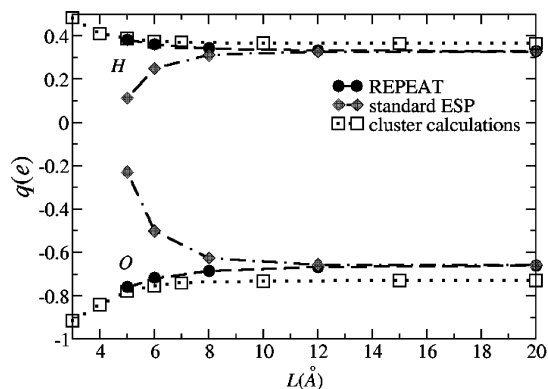*J. Chem. Theory Comput., Vol. 5, No. 10, 2009* **2873**

Figure 3. Charges of the O and H atoms of water as a function of the length of the simulation box using various methods.

undoubtedly a result of the ill-defined reference state in the periodic DFT calculations. As previously reasoned, Figure 3 shows the REPEAT charges and those derived using the standard functional converge to the same value as the simulation box increases. This suggests that the corrections due to the presence of the δ_ϕ term in eq 4 tend to zero as the simulation cell size increases.

To examine if the polarization of the REPEAT charges occurs in a physically suitable manner, we have performed a series of cluster calculations whereby a central water molecule is surrounded by 124 water molecules. The surrounding water molecules are positioned such that they correspond to the first and second shell of periodic images to mimic the truly periodic VASP calculations. The CHELPG charges of the central water molecule resulting from these cluster calculations are also plotted in Figure 3. At large simulation cell sizes, the charges resulting from the cluster calculation and the REPEAT charges converge to slightly different values. This is to be expected because of the differences in calculations, such as the nature of the basis sets used. Importantly, at short box lengths, the degree of polarization of the REPEAT charges is in good agreement with that observed in the cluster calculations. Thus, if molecular ESP derived charges are considered suitable for use in the simulation of molecular systems, then the results shown in Table 2 and Figure 3 suggest that REPEAT charges are equally appropriate for the simulation of periodic systems.

3. Dependency on Exclusion Radii and Grid Point Density. In this section, we examine how robust the REPEAT charges are with respect to the grid point density and the size of the VDW radii used to exclude grid points from the fitting procedure. To shed light on how the REPEAT charges vary with respect to these factors, we have selected two different porous materials: sodalite and zinc silicate. The results are summarized in Table 3 and Figure 4.

In most periodic electronic structure packages, the electrostatic potential is evaluated on a regular real space grid that is used for the discrete fast Fourier transform (FFT). For the practical application of the REPEAT method, it is of interest to know whether a typical FFT grid used in these calculations is appropriate for calculating the charges or whether a much higher grid density is required. To examine this, we have performed plane wave DFT calculations on sodalite and zinc

Table 3. REPEAT Charges for Sodalite and Zinc Silicate at Varying Grid Point Densities

molecule	atom	charge[a] or grid details		
		fine grid	medium grid	coarse grid
sodalite	Si	+1.355	+1.352	+1.372
	O	−0.677	−0.676	−0.686
	grid mesh	$100 \times 100 \times 100$	$50 \times 50 \times 50$	$20 \times 20 \times 20$
	grid points[b] per atom	10 925	1 359	88
zinc silicate[b]	Zn	+1.033(3)	+1.030(5)	+1.027(3)
	O	−0.493(3)	−0.494(3)	−0.485(3)
	Si	+1.387(3)	+1.383(4)	+1.377(6)
	grid mesh	$120 \times 160 \times 80$	$60 \times 80 \times 40$	$30 \times 40 \times 20$
	grid points[b] per atom	6 215	778	98

[a] Average charge (in e) over symmetry equivalent atoms. For zinc silicate, the standard deviation is reported in parentheses, while for sodalite, the standard deviation was 2×10^{-6} e or smaller in all cases. [b] Number of valid grid points outside of the VDW radii of the atoms used in the ESP fit. [b] Only the charges of selected Si and O atoms are presented.

silicate using an 80 Ry plane wave cutoff with the CPMD package. For sodalite, which has a simple cubic unit cell with length $a = 8.965$ Å, the plane wave cutoff results in a relatively dense $100 \times 100 \times 100$ grid point mesh. For zinc silicate, which has an orthorhombic unit cell with lengths $a = 10.07$, $b = 14.05$, and $c = 7.07$ Å, an 80 Ry plane wave cutoff in CPMD gives a default mesh of $120 \times 160 \times 80$ grid points. REPEAT charges of sodalite and zinc silicate have been evaluated using these default grid point meshes and are provided in Table 3 under the 'fine grid' subheading. The number of valid grid points per atom that lie outside of the VDW radii of the atoms that is used for the ESP fit is also given in Table 3. We then evaluated the REPEAT charges from the same DFT calculation, but where the grid points were uniformly trimmed. For example for sodalite, the grid points were trimmed by a factor of 8 and 125 to give $50 \times 50 \times 50$ and $20 \times 20 \times 20$ grid point meshes, respectively. These charges are given in Table 3 under the 'medium grid' and 'coarse grid' subheadings, respectively. Table 3 shows that the REPEAT charges remain essentially unchanged despite a drastic reduction in the number of grid points used in the fits. We note that the crystal symmetry was not imposed on the atomic charges during the fitting procedure. In the case of sodalite, which only has one symmetry unique Si atom and one unique O atom in a 36 atom unit cell, the standard deviation in the fitted charges was less than 2×10^{-6} e in all cases. For zinc silicate, the standard deviation was larger than with sodalite, but no greater than 6×10^{-3} e in all cases.

For sodalite, a $20 \times 20 \times 20$ FFT mesh (the 'coarse grid') would be considered very sparse for any periodic DFT package we are familiar with and translates into only 88 fitting points per atom. Thus, we believe that the default FFT mesh settings used in most periodic electronic structure codes should be appropriate for deriving REPEAT charges.

Due to the lack of a unique definition of VDW radii, there should ideally be no significant oscillations in the REPEAT charges with modest variations in the radii used to exclude grid points from the fitting procedure. Shown in Figure 4 are the REPEAT charges of sodalite and zinc silicate plotted as a function of the factor used to uniformly scale the default VDW radii. In the scaling range of 0.8−1.3, the charges on
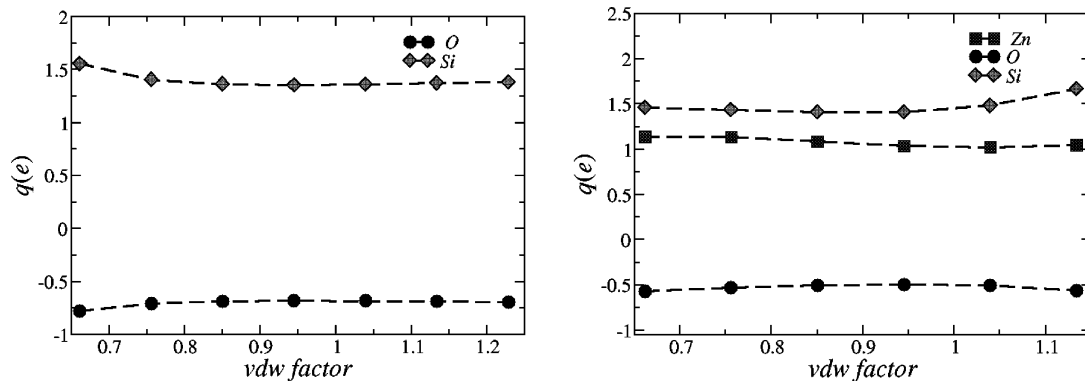
Figure 4. REPEAT charges for sodalite and zinc silicate under proportional scaling of the VDW radii used in the fitting procedure.

Si and O only differ by a maximum of 3.7% from the values derived using the default VDW radii. On an absolute scale, the O charges range from -0.703 to -0.678 e and for Si they range from $+1.406$ to $+1.356$ e. It is only when the VDW radii is rescaled by a factor of 0.7 that the charges begin to significantly deviate, differing by 14.7%. The REPEAT charges of zinc silicate are somewhat more sensitive to the VDW radii, with the charges deviating by more than 10% at scaling factors of 1.2 or greater and 0.7 or less. Such variations in the ESP fitted charges as a function of the VDW radii are related, in part, to alterations in the spatial geometry of the grid.

Our motivation for developing the REPEAT method was to generate charges for use in the classical simulation of microporous materials. Thus, we chose sodalite and zinc silicate to evaluate the method's sensitivity toward the VDW radii. We found that the total number of valid grid points could be drastically reduced without having a significant impact on the charges derived. As a result, we were curious as to how the method would perform when applied to densely packed solids, where the number of valid grid points available to perform the ESP fit is relatively small compared to microporous materials. For this purpose, we have examined the charges in zinc oxide, tin oxide, and cadmium telluride. Metal oxides like ZnO and SnO_2 enjoy a wide range of applications, ranging from gas sensors to transparent conductors in display technologies, while cadmium telluride has found important applications in solar cell technologies. VASP calculations of ZnO, SnO_2, and CdTe have been performed to generate an electrostatic potential for each compound. Using our default values for the VDW radii in these densely packed crystals resulted in mesh resolutions of approximately 10^3-10^4 grid points. Such numbers represent a total of at least 1 order of magnitude smaller than the typical grid sizes for microporous compounds, which have 3×10^5 points forming the grid. The variations in the charges as a function of the VDW scaling factor for zinc oxide, tin oxide, and cadmium telluride are given in Figure 5.

The REPEAT charges of oxygen in ZnO and SnO_2 appear to converge to the same approximate value of -1.25 e as a function of the VDW scaling. This is interesting since oxygen has a formal charge of -2 in both oxides. However, for zinc oxide, the convergence occurs at scaling factors greater than
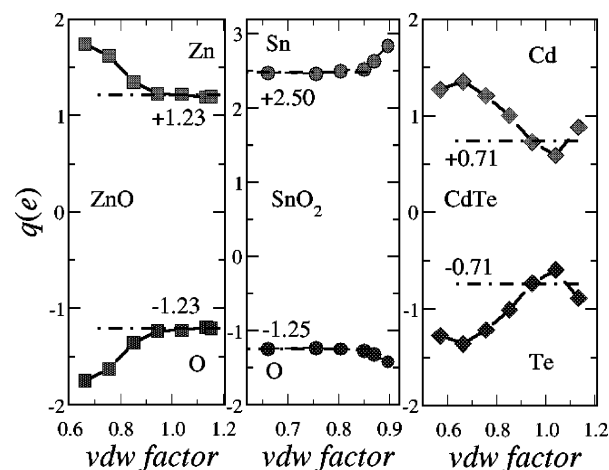

Figure 5. Comparison of the charges obtained for ZnO, SnO_2, and CdTe as a function of the VDW radii.

1, while it converges for values less than 1 for tin oxide. Another notable result is that the ESP charges for both the Zn and O atoms in ZnO are significantly greater in magnitude than those of the ESP charges of Zn and O in zinc silicate. This is consistent with the notion that the bonding in zinc silicate is more covalent, whereas the bonding in zinc oxide is considered more ionic in nature.

The results for CdTe are more difficult to analyze. As can be seen from Figure 5, the REPEAT charges oscillate and do not appear to converge as a function of the VDW radii used. This may be due to the fact that the charge is not as localized as in ZnO or SnO_2, whose bonding may be expected to be more ionic than in CdTe. This is supported by the fact that the charges in CdTe are notably smaller in magnitude than in ZnO, even though the formal charges are the same. Regardless of the underlying reasons for the behavior of the CdTe REPEAT charges, the results likely highlight the limitations of employing ESP derived charges for highly packaged systems.

4. Application to Porous Materials. Although one can foresee applications of the REPEAT method to various periodic systems, our motivation for developing the approach was our recent interest in simulating microporous and nanoporous materials, in particular MOFs. In this section, we apply the REPEAT method to two porous materials − a zeolite for which experimentally derived charges are available to compare with, and an MOF that has been the subject of

Electrostatic Potential Derived Atomic Charges

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2875**

a number of recent computational studies where different partial atomic charges used in the classical simulations are also available to compare with.

The ultimate goal of any theoretical study with practical applications in mind is to be able to reproduce/predict experimental (or experimentally derived) data. Although atomic charges are not a physical observable, the detailed analysis of X-ray diffraction data can allow for electronic density modeling and estimation of atomic charges, for example, via multipolar refinement. Coppens and others have developed a few procedures to extract partial atomic charges from experimental X-ray crystallographic data in this way.[49−51] In the context of molecular simulation, Pearlman and Kim have compared experimentally derived charges to those typically used in force fields for a molecular crystal of a nucleic acid constituent, 2′-deoxycytidine-5′-monophosphate.[52] For microporous materials, one of the works that we are aware of where experimental partial atomic charges have been derived is that of Ghermani et al.,[53] who examined natrolite. Although the charges should not be strictly the same, we thought it would be of interest to compare REPEAT charges of natrolite using the experimental geometry to those derived experimentally.

Natrolite is a naturally occurring zeolite whose structure has been studied since the 1930s when Pauling first examined it.[53] Natrolite ($Na_2Al_2Si_3O_{10}, 2H_2O$) has a unit cell containing a total of 184 atoms and possesses an orthorhombic crystal with a space group symmetry *Fdd*2. We have evaluated the electrostatic potential from a VASP calculation using the experimental lattice parameters and the atomic positions.[53] The tabulation of the electrostatic potential had a resolution of $N = 200 \times 200 \times 72$ mesh points. In the original work in which natrolite charges were derived from X-ray diffraction data, its authors had fixed the atomic charge of the sodium ions to be unity ($+1$). Aside from basic chemical intuition, the authors justified the sodium charges by the sharpness of the 4s scattering factor of the cited element within the diffraction pattern. Therefore, the charges of all atomic sites were computed in a reference state in which the charge of the sodium atom was always $+1$.

Shown in Table 4 are the unrestrained REPEAT charges determined for natrolite generated from a VASP and a CPMD calculation. To compare to the experimentally derived charges, we have imposed a strong quadratic RESP-like penalty function on the sodium atoms to restrain those charges to $+1$. We have also derived REPEAT charges where the Na atoms were allowed to freely fluctuate. The REPEAT charges are in general agreement with the experimentally derived charges. Perhaps the most notable difference is that the relative magnitude of the REPEAT charges on Si_1 and Si_2 are the opposite of what was determined experimentally. Although the charge was restrained to match the experimental charge on Na, the overall agreement between the experimental charges and the REPEAT charges, as quantified by Δ^2, was slightly reduced compared to the unrestrained REPEAT charges. Since Mulliken charges are still widely used in simulation of microporous materials, these charges derived for natrolite are also provided in Table 4. These charges were obtained from a CPMD calculation whereby

Table 4. Atomic Charges Determined for Natrolite Using Various Methodologies

atom[a]	X-ray derived[b]	REPEAT (Na = +1),[d]	REPEAT[c] VASP	REPEAT CPMD	Mulliken[e]
Si_1	+1.84(12)	+1.376	+1.370	+1.304	+1.541
Si_2	+1.65(10)	+1.722	+1.516	+1.451	+1.484
Al	+1.51(11)	+1.616	+1.435	+1.383	+1.165
O_1	−0.90(5)	−0.898	−0.851	−0.814	−0.602
O_2	−1.21(5)	−1.057	−0.895	−0.862	−0.629
O_3	−1.03(5)	−1.107	−0.989	−0.954	−0.619
O_4	−1.07(5)	−1.099	−0.981	−0.946	−0.615
O_5	−0.87(5)	−0.767	−0.730	−0.697	−0.598
Na	+1.0000	+1.000	+0.882	+0.862	+0.930
O_w	−0.59(3)	−1.002	−0.858	−0.807	−0.488
H_1	+0.24(3)	+0.472	+0.408	+0.383	+0.318
H_2	+0.36(3)	+0.432	+0.377	+0.350	+0.331
Δ^{2f}	–	0.501	0.489	0.609	1.156

[a] Atom numbering is taken from ref 53. [b] Ref 53 where the charge of Na is set to +1.0. [c] REPEAT charges derived from a VASP calculation of natrolite using the experimental geometry of ref 53. [d] REPEAT charges derived from VASP where the charge of Na is restrained to +1.0. [e] Mulliken charge derived from the CPMD calculation in which the wave functions expanded in terms of plane waves are projected[54] onto a single-ζ atom centered Slater-type functions. [f] Δ^2 is the sum of the differences between the experimentally derived reference charges and the calculated charges squared, $\Delta^2 = \Sigma(q_i^{\text{ref}} - q_i^{\text{calc}})^2$.

the plane wave orbitals were projected[54] onto a atom centered basis set composed of single-ζ Slater-type functions. The Mulliken charges are also in agreement with the experimentally derived charges, although the square difference, Δ^2, is approximately twice as large as the REPEAT charges. For reference, the REPEAT charges resulting from the CPMD calculation used to derive the Mulliken charges are also given in Table 4. Here it is notable that the REPEAT charges from CPMD and VASP (free Na charge) are very similar with a Δ^2 between one another of only 0.12.

One prototypical MOF material that has been the subject of several recent computational studies[12,23,32,55−59] is IR-MOF-1, also known as MOF-5.[8] IRMOF-1 has attracted significant attention as the parent compound of materials that have potential hydrogen storage applications.[60] The structure of IRMOF-1 consists of inorganic Zn_4O clusters that are each connected to four 1,4-benzenedicarboxylate (BDC) organic linkers in an octahedral array to form a porous cubic framework. Figure 6 shows a ball and stick model of the unit cell of IRMOF-1, which consists of 424 atoms.

A number of groups have examined various aspects of IRMOF-1 using classical potentials where the atomic charges have been primarily generated from fragment ESP calculations. Using Monte Carlo simulations, Sagara et al.[32] examined the H_2 binding in IRMOF-1 using CHELPG charges derived from a molecular fragment calculation of $Zn_4O(CO_2CH_3)_5-BDC-Zn_4O(CO_2CH_3)_5$ at the PBE/6-31+G(d) level of theory. Schmid and co-workers[12] studied the diffusion of benzene in IRMOF-1 with charges derived from fragment calculations[58] in a manner similar to Sagara's work but using the Merz−Kollman[61] ESP charge derivation scheme. Greathouse and co-workers have examined the interaction of water in IRMOF-1 with molecular dynamics simulations using charges that are based on the CVFF force field[62] that were then manually adjusted to more closely match the charges of Sagara. The charges used in these
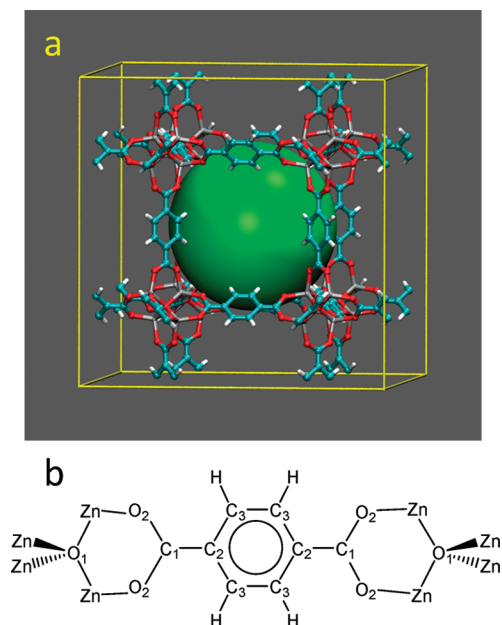
Figure 6. (a) Ball and stick representation of the unit cell of IRMOF-1 (blue = carbon, red = oxygen, white = hydrogen, gray = zinc). The green sphere represents the largest sphere that fits in the unit cell's cavity without touching the VDW's surface of the atoms. (b) Numbering of the unique atoms in IRMOF-1.

Table 5. Atomic Charges Determined for IRMOF-1

atom[a]	REPEAT	Sagara[b]	Schmid[c]	Greathouse[d]
Zn	+1.28	+1.31	+1.26	+1.20
O_1	−1.57	−1.79	−1.44	−1.20
O_2	−0.61	−0.63	−0.67	−0.60
C_1	+0.52	+0.62	+0.68	+0.60
C_2	+0.14	+0.05	+0.06	+0.00
C_3	−0.18	−0.12	−0.16	−0.10
H	+0.17	+0.12	+0.16	+0.10
Δ^{2e}	−	0.074	0.053	0.181

[a] Labeling of the atoms is given in Figure 6b. [b] Ref 32. [c] Ref 12. [d] Ref 56. [e] Δ^2 is defined in Table 2 with respect to the REPEAT charges.

simulation studies are given in Table 5 with the atom labeling provided in Figure 6b.

We have determined the REPEAT charges of IRMOF-1 resulting from the DFT calculation of the full periodic system using the CPMD package. These are given in Table 5 and show significant similarity to the charges reported by Sagara and Schmid resulting from the molecular fragment calculations and also to the charges of Greathouse that were manually adjusted to be similar to those derived by Sagara. These results further show that the REPEAT method produces sensible partial atomic charges for periodic systems. On the other hand, the results also put into question the need for an ESP charge derivation method for periodic systems, if molecular fragment calculations are adequate. We note that the suitability of ESP fragment calculations to ascertain partial charges in this particular compound relies on the fact that a chemically sensible and neutral moiety can be easily extracted from the framework. However, the extraction of fragments from a periodic solid can sometimes have non-trivial complications. For example, the need to 'cap' the

fragment to satisfy unfilled valences can, among other things, compete with the ideal of a charge neutral fragment. One such example within the MOF family occurs with zeolitic imidazolate framework (ZIF) materials that have recently been shown to have remarkably high CO_2 storage capabilities.[5,7] ZIFs are composed of Zn^{2+} ions tetrahedrally coordinated to imidazolate anions, Im^-, in a 1:2 ratio to form structural topologies identical to those of zeolites. Extraction of chemically sensible fragment with one Zn ion, $[ZnIm_4]^{2-}$, leaves the imidazolate anions under coordinated. Adding another layer of Zn ions gives $[Zn_5Im_4]^{6+}$, unfortunately, a rather large net fragment charge. Alternatively, the REPEAT approach offers a simple yet robust approach to deriving atomic charges in periodic systems such that the complications associated with extracting a sensible molecular fragment do not have to be contended with because they are determined from a true periodic calculation. This means that minimal user intervention is required to generate the REPEAT charges, which is necessary for the automation of the charge derivation procedure. Such automation would be very useful in high throughput computational studies of nanoporous materials.

V. Conclusions

A simple and robust method to derive ESP charges in periodic systems is introduced that we have termed the REPEAT method. To the best of our knowledge, the REPEAT method is the first ESP fitting technique that properly circumvents the fundamental issue of the ill-defined offset in the electrostatic potential within periodic electronic structure calculations. By modifying the conventional error functional used in molecular ESP charge derivation methods, we have developed an approach that can be equally applied to molecular systems as well as to periodic systems such as zeolites. A formal analytical proof is provided showing that the conventional error functional used in molecular ESP calculations is a limiting case of the REPEAT method for an infinitely large simulation box. The methodology also implicitly includes corrections to the dipole and the higher multipole moments when evaluating charges of isolated molecules with a finite-size periodic simulation cell.

Validation of the new methodology is demonstrated with a variety of tests. First it is shown that when the conventional ESP functional is applied to a continuous periodic solid, the fitted charges obtained with different periodic electronic structure packages differ drastically. In some cases, the charges obtained are 'unphysical', for example +10.5 e for Si in sodalite. In contrast, the REPEAT method generates charges that are both physically reasonable and very similar between different electronic structure codes. Tests on isolated molecules were also performed. The REPEAT charges, resulting from periodic calculations of 'isolated' molecules, were found to be in good agreement with the CHELPG and the RESP charges resulting from nonperiodic molecular DFT calculations.

The stability of the REPEAT charges, with respect to the VDW radii used to exclude fitting grid points, as well as the grid point density was examined. For microporous materials, such as sodalite, the REPEAT charges were found to be very

Electrostatic Potential Derived Atomic Charges

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2877**

stable upon variations of the aforementioned parameters, in particular the grid point density. To examine the limitations of the REPEAT method, several nonporous ionic solids were studied, specifically SnO_2, CdTe, and ZnO. The REPEAT charges of these systems were found to be chemically intuitive, however they did vary significantly with the VDW radii in some cases.

Although other applications can be foreseen, the original motivation for developing the REPEAT method was to generate partial atomic charges for the simulation of nanoporous materials, such as zeolites and metal organic framework materials. To demonstrate the applicability of the method to these systems, REPEAT charges were generated for the zeolite natrolite and the MOF IRMOF-1. Partial atomic charges have been determined from experimental X-ray diffraction data for natrolite and were found to be in good agreement with the REPEAT charges, although a direct correspondence is not expected. The second material, IR-MOF-1, has been the subject of a number of computational studies where charges have been derived from molecular ESP calculations on fragments extracted from the periodic framework. The REPEAT charges were in excellent agreement with these charges. Extracting chemically sensible fragments from a periodic solid to perform molecular ESP charge derivation calculations is not always trivial and minimally requires some user intervention or manual adjustment. Therefore, the REPEAT method offers a straightforward and robust approach to deriving ESP charges for the simulation periodic systems that is amenable to automation.

References

(1) *Handbook of Zeolite Science and Technology* Auerbach, S. M.; Carrado, K. A.; Dutta, P. K., Eds.; Taylor & Francis, Inc.: New York, 2003.

(2) James, S. L. *Chem. Soc. Rev.* **2003**, *32*, 276.

(3) Yan, Y.; Lin, X.; Yang, S. H.; Blake, A. J.; Dailly, A.; Champness, N. R.; Hubberstey, P.; Schroder, M. *Chem. Commun.* **2009**, 1025.

(4) Eddaoudi, M.; Kim, J.; Rosi, N.; Vodak, D.; Wachter, J.; O'Keeffe, M.; Yaghi, O. M. *Science* **2002**, *295*, 469.

(5) Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O'Keeffe, M.; Yaghi, O. M. *Science* **2008**, *319*, 939.

(6) Dinca, M.; Long, J. R. *Angew. Chem., Int. Ed.* **2008**, *47*, 6766.

(7) Banerjee, R.; Furukawa, H.; Britt, D.; Knobler, C.; O'Keeffe, M.; Yaghi, O. M. *J. Am. Chem. Soc.* **2009**, *131*, 3875.

(8) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. *Nature* **1999**, *402*, 276.

(9) Chae, H. K.; Siberio-Perez, D. Y.; Kim, J.; Go, Y.; Eddaoudi, M.; Matzger, A. J.; O'Keeffe, M.; Yaghi, O. M. *Nature* **2004**, *427*, 523.

(10) Smit, B.; Siepmann, J. I. *Science* **1994**, *264*, 1118.

(11) *Computer Modelling of Microporous Materials*; Catlow, C. R. A.; Santen, R. A. v.; Smit, B., Eds.; Elsevier Academic Press: Amsterdam, The Netherlands, 2004.

(12) Amirjalayer, S.; Tafipolsky, M.; Schmid, R. *Angew. Chem., Int. Ed.* **2007**, *46*, 463.

(13) Salles, F.; Ghoufi, A.; Maurin, G.; Bell, R. G.; Mellot-Draznieks, C.; Ferey, G. *Angew. Chem., Int. Ed.* **2008**, *47*, 8487.

(14) Han, S. S.; Deng, W. Q.; Goddard, W. A. *Angew. Chem., Int. Ed.* **2007**, *46*, 6289.

(15) Han, S. S.; Goddard, W. A. *J. Am. Chem. Soc.* **2007**, *129*, 8422.

(16) Leach, A. R. *Molecular Modelling Principles and Applications*; 2nd ed.; Pearson Education Limited: Harlow, England, 2001, pp. 189−191.

(17) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(18) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(19) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300.

(20) Liu, D.; Zheng, C.; Yang, Q.; Zhong, C. *J. Phys. Chem. C* **2009**, *113*, 5004.

(21) Yang, Q. Y.; Zhong, C. L. *J. Phys. Chem. B* **2006**, *110*, 17776.

(22) Dubbeldam, D.; Walton, K. S.; Ellis, D. E.; Snurr, R. Q. *Angew. Chem., Int. Ed.* **2007**, *46*, 4496.

(23) Greathouse, J. A.; Allendorf, M. D. *J. Phys. Chem. C* **2008**, *112*, 5795.

(24) Blöchl, P. E. *J. Chem. Phys.* **1995**, *103*, 7422.

(25) Blöchl has developed a method24 to derive atomic partial charges from the total electron density of periodic plane wave calculations. However, the method is only valid for the calculation of isolated molecules or clusters when utilizing the supercell approximation in a periodic calculation.

(26) Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods* Cambridge University Press: Cambridge, U.K., 2004, 499−502.

(27) See Computational Details, Section III, for full details.

(28) Ewald, P. *Ann. Phys.* **1921**, *64*, 253.

(29) de Leeuw, S. W.; Perram, J. W.; Smith, E. R. *Proc. R. Soc. Lond. A* **1980**, *373*, 27.

(30) Smith, P. E.; Pettitt, B. M. *Comput. Phys. Commun.* **1995**, *91*, 339.

(31) *CPMD User's Manual* , Version 3.13.2, http://www.cpmd.org. Accessed August 6, 2009.

(32) Sagara, T.; Klassen, J.; Ganz, E. *J. Chem. Phys.* **2004**, *121*, 12543.

(33) Iczkowsky, R. P.; Margrave, J. L. *J. Am. Chem. Soc.* **1961**, *83*, 3547.

(34) Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.

(35) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.

(36) Kresse, G.; Furthmüller, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169.

(37) *CPMD*, Version 3.13.2; IBM Research Division and MPI für Festkörperforschung Stuttgart: 2008.

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

(39) Soler, J. M.; Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Ordejon, P.; Sanchez-Portal, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745.

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

(41) Blöchl, P. E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953.

(42) Kresse, G.; Joubert, D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 1758.

(43) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 1703.

(44) Case, D. A.; Darden, T. A.; III, T. E. C.; 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.

(45) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.

(46) Mabilia, M.; Pearlstein, R. A.; Hopfinger, A. J. *J. Am. Chem. Soc.* **1987**, *109*, 7960.

(47) Nicholas, J. B.; Hopfinger, A. J.; Trouw, F. R.; Iton, L. E. *J. Am. Chem. Soc.* **1991**, *113*, 4792.

(48) Although Gaussian 03 can perform periodic calculations, it cannot print out the electrostatic potential for periodic calculations.

(49) Hirshfeld, F. L. *Acta Crystallogr., Sect. B: Struct. Sci.* **1971**, *27*, 769.

(50) Hansen, N. K.; Coppens, P. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 909.

(51) Coppens, P.; Guru Row, T. N.; Leung, P.; Stevens, E. D.; Becker, P. J.; Yang, Y. W. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1979**, *35*, 63.

(52) Pearlman, D. A.; Kim, S.-H. *Biopolymers* **1985**, *24*, 327.

(53) Ghermani, N. E.; Lecomte, C.; Dusausoy, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *53*, 5231.

(54) Sanchez-Portal, D.; Artacho, E.; Soler, J. M. *Solid State Commun.* **1995**, *95*, 685.

(55) Greathouse, J. A.; Allendorf, M. D. *J. Am. Chem. Soc.* **2006**, *128*, 10678.

(56) Greathouse, J. A.; Allendorf, M. D. *J. Am. Chem. Soc.* **2006**, *128*, 13312.

(57) Mulder, F. M.; Dingemans, T. J.; Wagemaker, M.; Kearley, G. *J. Chem. Phys.* **2005**, *317*, 113.

(58) Tafipolsky, M.; Amirjalayer, S.; Schmid, R. *J. Comput. Chem.* **2007**, *28*, 1169.

(59) Kim, D.; Kim, J.; Jung, D. H.; Lee, T. B.; Choi, S. B.; Yoon, J. H.; Kim, J.; Choi, K.; Choi, S.-H. *Catal. Today* **2007**, *120*, 317.

(60) Rosi, N. L.; Eckert, J.; Eddaoudi, M.; Vodak, D. T.; Kim, J.; O'Keeffe, M.; Yaghi, O. M. *Science* **2003**, *300*, 1127.

(61) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.

(62) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. *Proteins: Struc. Funct. Genet.* **1988**, *4*, 31.

JCTC Journal of Chemical Theory and Computation

Quantum Mechanical Quantification of Weakly Interacting Complexes of Peptides with Single-Walled Carbon Nanotubes

Wenjie Fan,[†] Jun Zeng,[‡,§] and Ruiqin Zhang*[,†]

Centre of Super-Diamond and Advanced Films (COSDAF) and Department of Physics and Materials Science, City University of Hong Kong, Hong Kong SAR, China, MedChemSoft Solutions, P.O. Box 5143, Wantirna South, VIC 3152, Australia, and College of Chemistry, Sichuan University, Chengdu 610064, China

Abstract: We investigated the binding nature of three peptides (inactive NB1 and active B1 and B3) to single-walled carbon nanotubes (SWCNTs) using a density functional tight-binding (DFTB) method with an empirical van der Waals force correction. We show that the three peptides could be spontaneously adsorbed to the carbon nanotube (CNT) surface through $\pi-\pi$ and/or H$-\pi$ stacking at physisorption distances and the geometric and π-electronic structures of SWCNTs remain basically undamaged upon the adsorption. We also investigated the diameter and chirality dependence of binding energies. The calculated results are consistent with experimental observation, and we found that aromatic residues, such as His and Trp, are the keys in determining peptide/CNT binding. In addition, our calculations predict that noncovalent modification of SWCNTs by the active peptides might increase the electron transfer capabilities of SWCNTs.

1. Introduction

Carbon nanotubes (CNTs), discovered in 1991 by Iijima,[1] have many novel electrical and mechanical properties, such as high electrical conductivity, excellent stiffness against bending, and high tensile strength.[2] Single-walled carbon nanotubes (SWCNTs) have potential biological applications ranging from biomedical sensors to drug delivery.[3−5] But such biological applications have, so far, been limited owing to two major obstacles: hydrophobicity and conformational heterogeneity. Although the solubility in water can be improved by chemically modifying the SWCNTs through covalent bonding of various functional groups to the nanotubes,[6,7] these modifications can perturb the intrinsic properties of SWCNTs, such as electrical properties. As a result, alternative approaches using the noncovalent adsorp-

tion of surfactants,[8] polymers,[9] and biomolecules[10−12] to solubilize the SWCNTs have been proposed and tested.

In recent years, much research attention has focused on the design and utilization of polypeptide/CNT complexes because of their functionality in biological systems.[13−18] Encircling CNTs with peptides has two major advantages:[15] (1) the peptides will not dissociate from the CNTs thereby providing extremely stable CNT dispersions and (2) this method enables the diameter selective separation of SWCNTs,[15] since nanotubes with a specific diameter are preferentially circled by a given length of peptides. For example, to improve CNT solubility in water as well as biocompatibility, Zorbas et al. developed a family of amphiphilic helical peptides that noncovalently bind and solubilize SWCNTs in water, yielding unbundled and individual SWCNTs.[14] Later they devised a novel way to coat SWCNTs with reversible cyclic peptides to separate them based on differences in diameters.[15]

Despite these exciting experimental observations, many outstanding questions remain regarding the affinity of selected peptide/SWCNTs complexes. For example, how

---

* Corresponding author. E-mail: aprqz@cityu.edu.hk.
† City University of Hong Kong.
‡ MedChemSoft Solutions.
§ Sichuan University.

does the type of carbon-based nanomaterial differ in the selection for peptide sequence? Do peptide sequences selected for one kind of carbon-related nanomaterials preserve the high attraction for other forms? As an initial step in exploring these critical questions, we investigated the binding feature between SWCNTs and three representative peptides, namely, inactive peptide NB1 and active peptides B1 and B3. The present research is a theoretical contribution to the binding features of biomolecule/nanomaterial complexes in the gas phase.

2. Computational Details

Weak interactions exist widely and play a significant role in determining and stabilizing the three-dimensional structures of proteins. In proteins, the weak interactions among aromatic side chains, backbone amides, carboxyl groups, aliphatic hydrogen atoms, and other parts exist widely. Individually, these interactions might be weak but when combined they have a large influence on the conformational stability of a protein.[19,20] For example, the arene−arene and arene−COO interactions were reported to complicate the measurement of the UV absorption of residues.[21]

The peptides adsorbed on the surface of SWCNTs might be a proto-typical weakly interacting complex that involves π bonding; this remains a challenging issue for theoretical studies because these systems cannot be described correctly by conventional DFT or methods based on classical interatomic potentials. MP2 and CCSD theories are used as standard methods to consider the dispersion force,[22,23] but for large systems of biological interest, such methods are not applicable. In this study, we used a computationally efficient approximation to density functional theory, the self-consistent charge density functional tight-binding (SCC-DFTB) scheme, complemented by the empirical London dispersion energy term (acronym DFTB-D) to study the energy and geometry structure of the three peptides in relation to the CNTs. The SCC-DFTB model has been derived from a second-order expansion of the DFT total energy functional with respect to charge density fluctuations; at the same time, the Hamiltonian matrix elements are calculated with a two-center approximation. These are then tabulated together with the overlap matrix elements with respect to interatomic distance. A comprehensive description of the method can be found in the literature.[24,25] The van der Waals interaction has been described with an empirical dispersion term, consisting of an R^{-6}, term added to the SCC-DFTB total energy.[26,27] The SCC-DFTB-D has been successfully applied to investigating the energies and structures of biomolecules.

In this work, we considered SWCNTs with different chiralities, saturated at the ends by hydrogen atoms. The CNTs we studied are armchair (n, n) $(n = 3−6)$, zigzag (8,0) tube, and chiral (6,3) tube. The tube length is around 33 Å, which is long enough to eliminate the effect of the terminal hydrogen atoms. We investigated three representative peptides, including the peptide NB1 of sequence LPPSNAS-VADYS, the peptide B1 of sequence HWKHPWGAWDTL, and the peptide B3 of sequence HWSAWWIRSNQS. Peptides NB1 and B1 were initially constructed as fully extended

structures. For B3, previous CD experiments have demonstrated that the free B3 peptide adopts a helical conformation at a neutral pH value and the binding to SWCNTs does not affect its helical conformation.[18] Therefore, in this work, we considered the initial structure of peptide B3 as a helical conformation. We first positioned the arene clusters in these peptides approximately parallel to the sidewall of the SWCNTs, with part of them making direct contact, followed by full geometric optimizations using conjugate gradient algorithm until the residual forces were below 2×10^{-4} au; we set the charge convergence criterion to be 10^{-5} electrons.

To provide a better understanding of the electronic properties of the peptide/CNT complexes, we investigated the density of states (DOS) of the active complexes. The DOS was calculated using the SIESTA method.[28,29] We used the Ceperley−Alder version of the LDA[30] for the electron exchange and correlation as well as optimized Troullier−Martins pseudopotentials[31] for the atomic cores in this calculation. We performed the computations of this part using a supercell approach and periodic-boundary conditions. We adopted a lateral separation of 35 Å between the tube centers, which was large enough to eliminate the interaction between the neighboring tubes. We then used two Monkhorst−Pack k points for the Brillouin zone integration along the tube axis. A cutoff of 150 Ry for the grid integration was utilized, representing the charge density. Although different models might result in different optimized structures and, thus, different electronic properties, this effect can be eliminated if the tube length is long enough to eliminate the interaction between the peptide and the hydrogen atoms at the tube ends and the lateral separation between tube centers is large enough to eliminate the interaction between the neighboring tubes. In our work, the binding energies for the cluster and the periodic models were respectively almost the same, as were the optimized structures for the interacting parts for both models.

3. Results and Discussion

3.1. Structures and Binding of the Peptide/CNT Complexes. Figures 1a−c show the optimized structures for the three peptides. NB1 involves only one arene (Tyr) at the 11th position. B1 in which an extended conformation is adopted involves five arenes: three tryptophans at the second, sixth, and ninth positions and two histidines at the first and fourth positions, respectively. For the helical peptide B3, the structure is relatively compact as a result of the intramolecular hydrogen bonds. One representative intramolecular hydrogen bond is formed between C=O of the Trp2 and N−H of the Trp6 positions with N−O and H−O distances of 2.91 and 1.90 Å, respectively.

Figure 2 shows the optimized geometries of the three peptides/(5,5) CNT complexes. We calculated the binding energies of the studied systems with the following equation: $E_{\mathrm{B}} = E_{\mathrm{(peptide/SWCNT)}} - E_{\mathrm{(SWCNT)}} - E_{\mathrm{(peptide)}}$, where $E_{\mathrm{(SWCNT)}}$ and $E_{\mathrm{(peptide)}}$ are the total energy of the isolated pristine tube and peptide, respectively, and $E_{\mathrm{(peptide/SWCNT)}}$ is the total energy of the peptide/SWCNT complex.
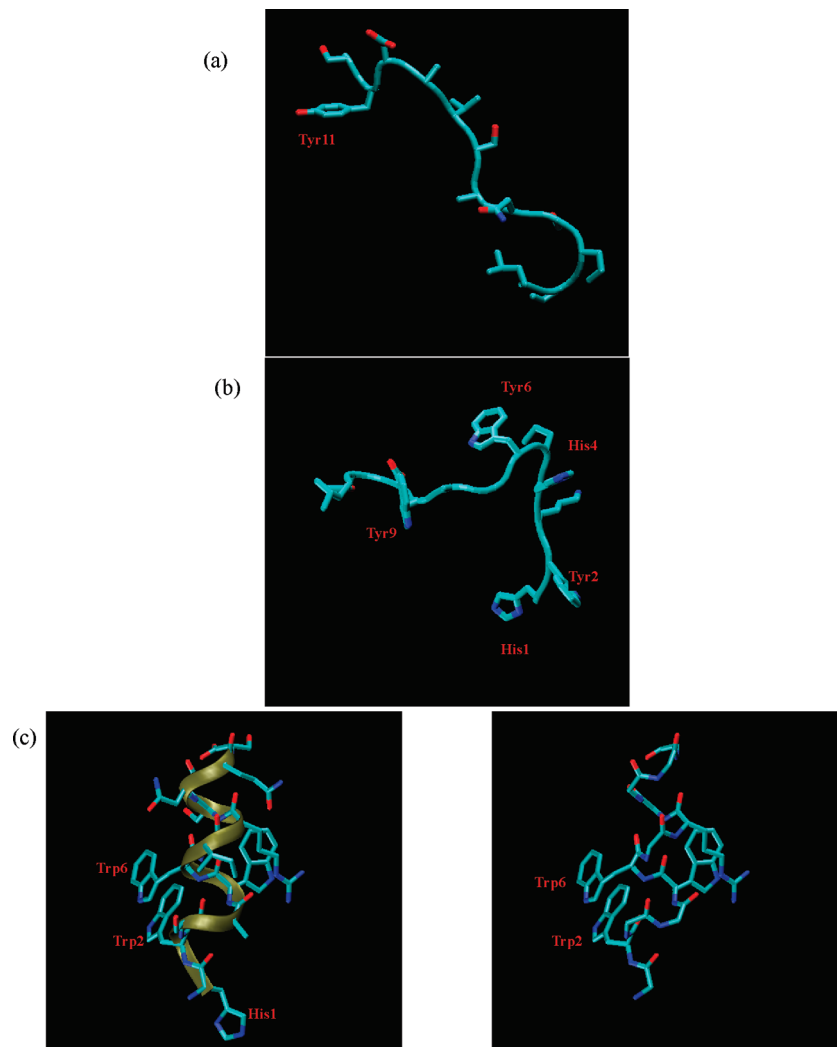
Figure 1. Optimized structures for (a) peptide NB1, (b) peptide B1, and (c) peptide B3. The important aromatic residues are labeled in red.

Upon adsorption to the CNTs, the three peptides show different binding features. For the inactive peptide NB1 (shown in Figure 2a), the contact area of peptide NB1 with the (5,5) CNT surface is located at the aromatic ring of Tyr11. The phenol ring is adsorbed to the CNT surface in a parallel manner by a favorite "bridge site" with a nearest C−C distance of 3.00 Å. The binding energy is calculated to be −0.71 eV. Since little contact occurs between the CNT and the other parts in NB1, the geometry of peptide NB1 undergoes little change upon adsorption.

The active peptide B1 wraps around the (5,5) CNT surface, as shown in Figure 2b. Compared with isolated B1 in Figure 1b, most of the peptide interacts with the CNT surface, and the adsorbed B1 shows obvious conformational changes to facilitate maximum interaction with the surface. First, the aromatic ring of the N-terminal residue His1 is turned parallel to the tube surface, with the nearest distance between the C−C around 3.14 Å. Second, of all three Trp residues, Trp2 and Trp6 adopt a favorable parallel orientation to the CNT surfaces. However, Trp9 is tilted to the CNT surface due to the steric effect. For the Trp2 ring that is favorably bonded to the CNT surface by $\pi-\pi$ stacking, the nearest distance between the C−C is around 3.02 Å. While the His1 is close

to the tube surface, the His4 is perpendicular to the tube. Since the CNT is circled by the peptide B1, the contact area is much larger than that of NB1/CNT. Third, several XH−π (X = C or N) interactions occur between the peptide B1 and the CNT surface. While a NH−π interaction is found between the backbone amide of Trp2 and the CNT surface with a N−C distance of 3.25 Å and a H−C distance of 2.28 Å, respectively, a weaker CH−π interaction is identified between the Cα of Lys3 and the CNT surface, with a shortest C−C distance of 3.48 Å and a H−C distance of 2.47 Å, respectively. The overall binding energy is calculated to be −2.85 eV.

For active peptide B3 with a helical conformation, the binding energy to the (5,5) CNT is predicted to be −2.22 eV. Figure 2c shows the adsorbed peptide on the CNT surface. Compared with the isolated peptide, the adsorption has minor effect on the peptide's helical conformation. Similar to active peptide B1, peptide B3 binds to the CNT surface via the favorable $\pi-\pi$ stacking between the CNT surface and the aromatic residues His1, Trp2, and Trp5, with the corresponding N−C and C−C distances calculated to be 3.21 and 3.05 Å, 3.34 and 2.94 Å, and 3.22 and 3.06 Å, respectively. Moreover, the $\pi-\pi$ stacking clusters Trp2/Trp6
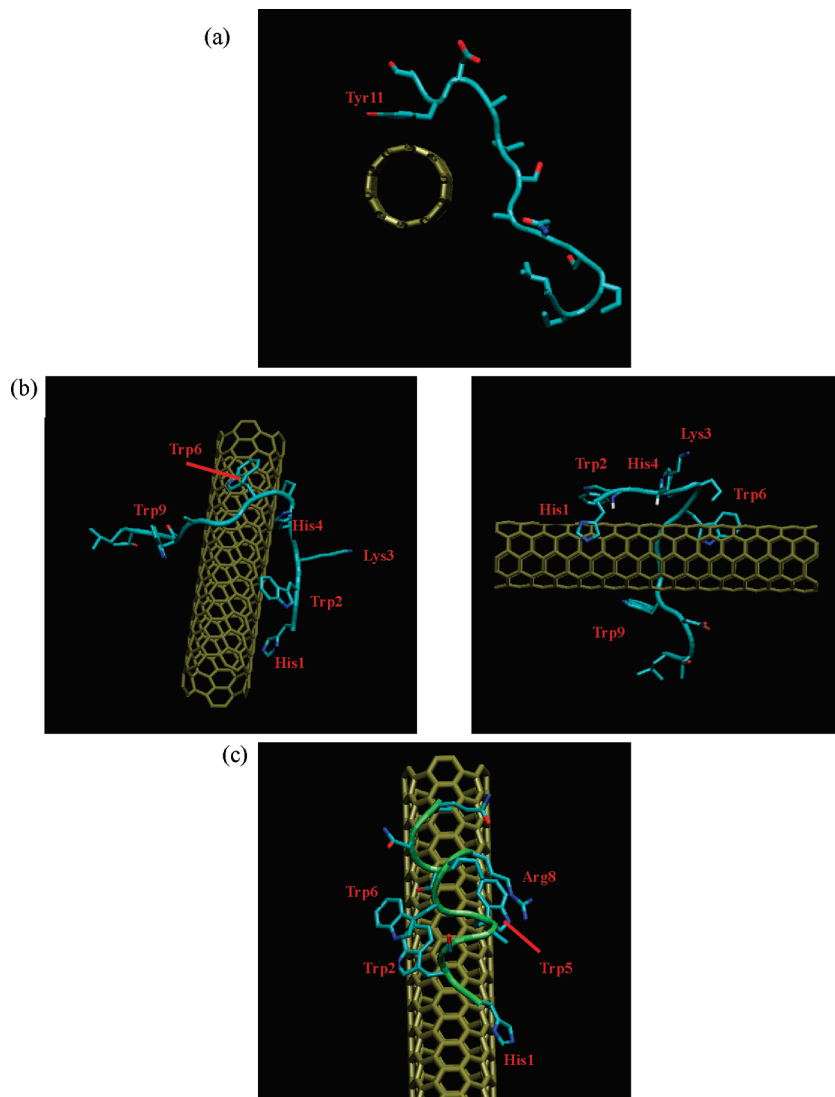
Figure 2. Illustration of the peptide/(5,5) CNT interactions: (a) inactive peptide NB1, (b) the active peptide B1, and (c) the active peptide B3. The $\pi-\pi$ stacking and XH$-\pi$ (X = C and N) interactions are displayed in the left and right panels for the (b) B1/CNT complex, respectively. Key residues are labeled in red.

and Trp5/Arg8 found in the isolated peptide (Figure 1c) are well preserved in the B3/CNT complex, indicating that the $\pi-\pi$ interactions contribute not only to the binding activities of the peptide but also to the stabilization of the helical conformation. The preservation of the helical conformation upon peptide B3 adsorption to the side wall of SWCNT is consistent with the experimental observation obtained using CD spectroscopy.[18] Our results demonstrate that the geometric structure of CNT remains almost unchanged after the adsorption of peptides.

3.2. Diameter and Chirality Dependence of the Binding Energy for the Peptide/CNT Complexes. Figure 3 shows the diameter dependence of the binding energies of active peptide B1 to the (n, n) CNTs ($n = 3-6$) at a given tube length. Overall, the binding energies increase as the tube diameter grows larger, except at the beginning when $n = 3$ and 4; the peptide/(4,4) CNT binding is slightly higher than the binding of the peptide/(3,3) CNT. For the active B1 that encloses the CNTs, the contact area between the B1 and the CNT is large. When B1 is adsorbed to the CNT, the aromatic parts (such as His and Trp) in direct contact with the CNT
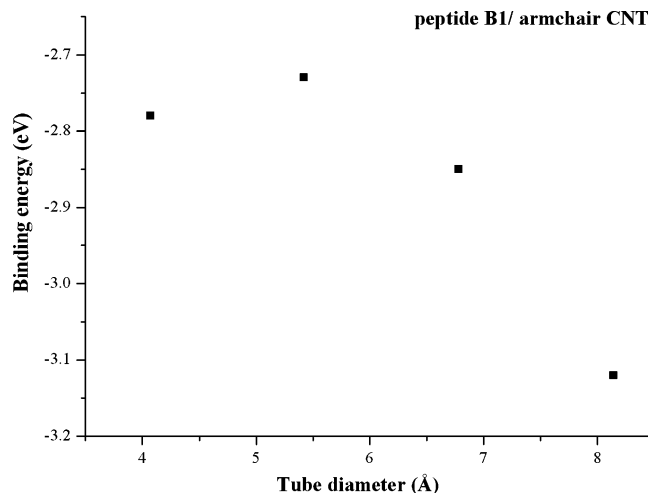

Figure 3. Diameter dependence of binding energies of B1 to the (n,n) CNT ($n = 3-6$) at a given tube length.

will undergo considerable conformation change to facilitate maximum intermolecular interactions with the CNT. This

Quantum Mechanical Quantification of Peptide Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2883**

Table 1. The Chirality Dependence of the Binding Energy for Peptide/CNT Complexes

	tube diameter (Å)	binding energies (eV)
NB1/(5,5) CNT	6.78	−0.71
NB1/(8,0) CNT	6.26	−0.67
NB1/(6,3) CNT	6.23	−0.64
B1/(5,5) CNT	6.78	−2.85
B1/(8,0) CNT	6.26	−2.82
B1/(6,3) CNT	6.23	−2.69
B3/(5,5) CNT	6.78	−2.22
B3/(8,0) CNT	6.26	−2.21
B3/(6,3) CNT	6.23	−1.97

will result in a change of B1 geometry arising from intramolecular interactions. In addition, the geometric change in peptide B1 will affect the intermolecular interactions between the CNT surface and peptide B1. This chain interaction will finally change the basic structure of the free peptide B1. As mentioned above, the interactions between B1 and CNT involve $\pi-\pi$, CH$-\pi$, and NH$-\pi$ interactions, and the competition of these weak interactions will determine the final equilibrium structures and binding energy. Therefore, we cannot simply assume that as the contact area increases, the binding energies will increase correspondingly, just as we found in 1-pyrenebutanoic acid, succinimidyl ester (PSE)/CNT.[32] We did not observe any linear relationship between the binding energy and the tube diameter for any of the three peptide/CNT complexes considered here.

We also studied the chirality dependence of the binding energy for peptides/CNT and considered three tubes with similar diameters: (5,5), (8,0), and (6,3) tubes. Table 1 gives the results. We found that, although the chirality dependence of the binding energies is not obvious, generally as the tube diameter increases, the binding energy will also increase. Of the three tubes considered here, the armchair (5,5) tube, with the largest tube diameter, has the largest binding energy, while the chiral (6,3) tube, with the shortest diameter, demonstrates the smallest binding energy.

3.3. Molecular Orbitals of the Peptide/CNT Complexes. Figure 4 presents the isosurfaces of the selected frontier orbitals for peptide B3/(5,5) CNT, pristine (5,5) CNT, and single B3 at a 0.01 au isovalue. For isolated B3, the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) locate on the Arg and Trp2, respectively. After adsorption to the (5,5) CNT system, both the HOMO and LUMO locate on the (5,5) tube, and no molecular orbital overlapping occurs between the B3 and the (5,5) CNT. We conclude that the noncovalent $\pi-\pi$ stacking of B3 to CNTs does not damage the π-conjugation electronic structures of the CNTs. Our calculations show that after adsorption of B3 on the sidewall, the HOMO$-$LUMO gap for the CNT remains almost unchanged. These results suggest that the π-electronic properties of the CNT are preserved after noncovalent adsorption of B3, which is similar to the features for simple planar organic molecules adsorbed on CNTs.[33,34] We find similar orbital features for the NB1/(5,5) CNT and the B1/(5,5) CNT.

3.4. Electronic Properties. We sought further understanding of the electronic properties of peptide/CNT interactions by calculating the DOS of the peptide/CNT complexes.
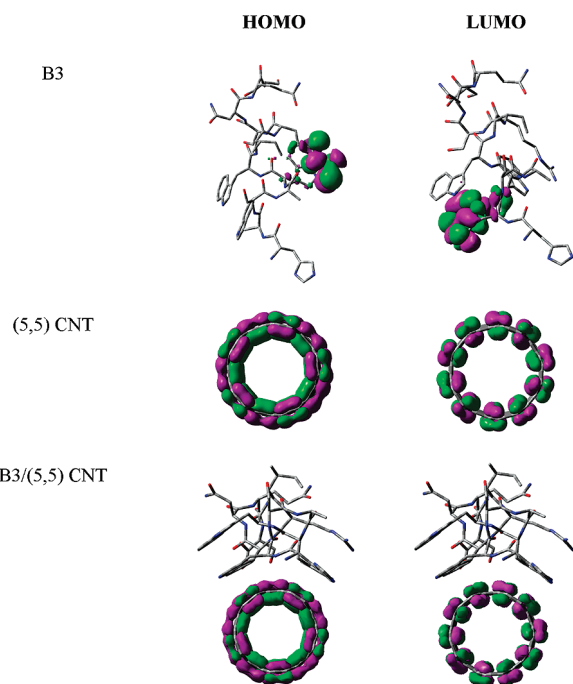
Figure 4. Isosurfaces of the selected frontier orbitals of the B3/(5,5) CNT complex at the ground state. The isovalue is 0.01 au.
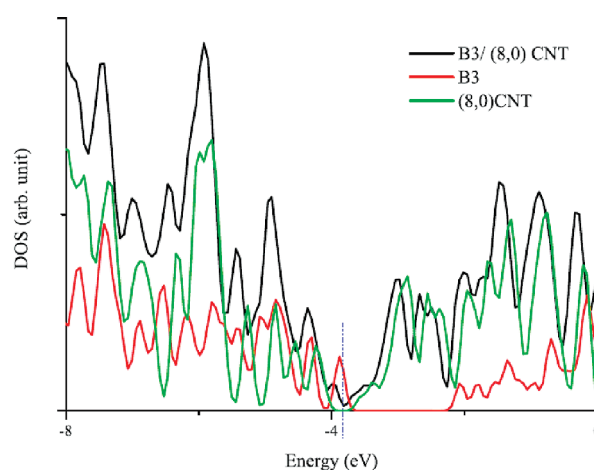
Figure 5. Total density of states for B3/(8,0) CNT.

We considered the metallic (5,5) CNT and the semiconducting (8,0) CNT to study the features of the DOS. Figure 5 shows the total DOS of the B3/(8,0) CNT. For the B3/metallic (5,5) CNT (not shown here), we find that the total DOS of the system preserves the DOS features of pristine (5,5) tube, and there isn't obvious new DOS induced near the Fermi level. But we find different binding features when B3 interacts with semiconducting (8,0) CNT. Compared with the pristine (8,0) CNT, the DOS of the B3/(8,0) CNT shows new states near the Fermi level, which were contributed by the peptide B3. Still, the nature of physisorption is shown since the total DOS preserves the features of the DOS of pristine (8,0) CNT. The Fermi energy level (E_f) of the semiconducting (8,0) CNT was predicted to be −3.86 eV, while for the B3/(8,0) CNT, the E_f was almost unchanged. We find new DOS, composed mainly of the HOMO of B3 molecular orbitals, between the (8,0) CNT conduction and

2884 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Fan et al.

valence bands. Hence, the band gap of the system has sharply decreased from the pristine (8,0) tube 0.58 to 0.29 eV of the complexes. Similar finding of band gap reduction has been found in the (7,3) polyC−DNA complex compared to the free semiconducing (7,3) tube.[35] For the semiconducting (8,0) CNT, the band gap remains almost unchanged. For peptide B3, our calculations show that the band gap is enlarged slightly after adsorption. The finding of the new DOS formed between the (8,0) CNT conduction and valence bands is quite similar to our previous study of flavin adenine dinucleotide (FAD)/(10,0) CNT.[36] For the DOS of B1/(8,0) CNT (not shown here), we find similar features of the weak interaction: new states, composed mainly of the HOMO of the B1 molecular orbital, between the (8,0) CNT conduction and valence bands. Compared to pristine SWCNTs, the new states formed in the active peptides/(8,0) system may cause obvious increasing electron transfer abilities. A CV experiment is highly desirable to verify this observation.

We furthermore examined the charge transfer between the peptide B3 and the (8,0) CNT from an analysis of the Mulliken charges for the SCC-DFTB-D calculations. There is very small charge transfer (0.02e) observed from peptide B3 to (8, 0) tube. Again, the negligible charge transfer in this π−π bonded system demonstrates the weak interaction nature of the peptide and CNT complexes.

Based on the above analysis of the structures and properties of the peptide/CNT complexes, we show that the arene parts, such as His and Trp, play a significant role in the noncovalent binding of peptides to the CNT surface. Our theoretical results confirm the experimental finding[37] and also agree with previous theoretical reports.[38,39]

4. Conclusions

We show via quantum mechanical investigation that peptides (inactive NB1 and active B1 and B3) could be spontaneously attracted to the sidewall of CNTs through π−π and/or H−π stacking, which is at the physisorption distance. The competition of π−π and/or H−π stacking plays a key role in binding the peptides to the CNTs, thus, determining and stabilizing the binding of the peptide/CNT systems. The isosurfaces of the selected frontier orbitals show that the π-electronic structures of the CNTs remain basically undamaged. This study confirms the experimental finding[37] of the key role of His and Trp when binding to the CNT surface. Our calculations predict that the noncovalent modification of SWCNTs by the active peptides might increase the former's electron transfer capabilities. These results provide a better understanding of the binding between peptides and CNTs and, therefore, have potential applications for designing biofunctionalized CNTs as biosensors and drug delivery devices.

References

(1) Iijima, S. *Nature* **1991**, *354*, 56–58.

(2) Ajayan, P. M. *Chem. Rev.* **1999**, *99*, 1787–1799.

(3) Martin, C. R.; Kohli, P. *Nat. Rev. Drug Discovery* **2003**, *2*, 29–37.

(4) Li, J.; Ng, H. T.; Chen, H. *Methods Mol. Biol.* **2005**, *300*, 191–23.

(5) Contarino, M. R.; Sergi, M.; Harrington, A. E.; Lazareck, A.; Xu, J.; Chaiken, I. *J. Mol. Recognit.* **2006**, *19*, 363–371.

(6) Hirsch, A. *Angew. Chem., Int. Ed.* **2002**, *41*, 1853–1859.

(7) Huang, W.; Fernando, S.; Lin, Y.; Zhou, B.; Allard, L. F.; Sun, Y.-P. *Langmuir* **2003**, *19*, 7084–7088.

(8) O'Connell, M. J.; Bachilo, S. M.; Huffman, C. B.; Moore, V. C.; Strano, M. S.; Haroz, E. H.; Rialon, K. L.; Boul, P. J.; Noon, W. H.; Ma, J.; Hauge, R. H.; Weisman, R. B.; Smalley, R. E. *Science* **2002**, *297*, 593–596.

(9) Dalton, A. B.; Blau, W. J.; Chambers, G.; Coleman, J. N.; Henderson, K.; Lefrant, S.; McCarthy, B.; Stephan, C.; Byrne, H. J. *Synth. Met.* **2001**, *121*, 1217–1218.

(10) Zheng, M.; Jagota, A.; Semke, E. D.; Diner, B. A.; Mclean, R. S.; Lustig, S. R.; Richardson, R. E.; Tassi, N. G. *Nat. Mater.* **2003**, *2*, 338–342.

(11) Zheng, M.; Jagota, A.; Strano, M. S.; Santos, A. P.; Barone, P.; Chou, S. G.; Diner, B. A.; Dresselhaus, M. S.; Mclean, R. S.; Onoa, G. B.; Samsonidze, G. G.; Semke, E. D.; Usrey, M.; Walls, D. J. *Science* **2003**, *302*, 1545–548.

(12) Numata, M.; Asai, M.; Kaneko, K.; Bae, A.-H.; Hasegawa, T.; Sakurai, K.; Shinkai, S. *J. Am. Chem. Soc.* **2005**, *127*, 5875–5884.

(13) Dieckmann, G. R.; Dalton, A. B.; Johnson, P. A.; Razal, J.; Chen, J.; Giordano, G. M.; Muñoz, E.; Musselman, I. H.; Baughman, R. H.; Draper, R. K. *J. Am. Chem. Soc.* **2003**, *125*, 1770–1777.

(14) Zorbas, V.; Ortiz-Acevedo, A.; Dalton, A. B.; Yoshida, M. M.; Dieckmann, G. R.; Draper, R. K.; Baughman, R. H.; Jose-Yacaman, M.; Musselman, I. H. *J. Am. Chem. Soc.* **2004**, *126*, 7222–7227.

(15) Ortiz-Acevedo, A.; Xie, H.; Zorbas, V.; Sampson, W. M.; Dalton, A. B.; Baughman, R. H.; Draper, R. K.; Musselman, I. H.; Dieckmann, G. R. *J. Am. Chem. Soc.* **2005**, *127*, 9512–9517.

(16) Karajanagi, S. S.; Yang, H.; Asuri, P.; Sellitto, E.; Dordick, J. S.; Kane, R. S. *Langmuir* **2006**, *22*, 1392–1395.

(17) Pender, M. J.; Sowards, L. A.; Hartgerink, J. D.; Stone, M. O.; Naik, R. R. *Nano Lett.* **2006**, *6*, 40–44.

(18) Su, Z.; Leung, T.; Honek, J. F. *J. Phys. Chem. B* **2006**, *110*, 23623–23627.

(19) Burley, S. K.; Petsko, G. A. *Adv. Protein Chem.* **1988**, *39*, 125–189.

(20) Hatfield, M. P. D.; Palermo, N. Y.; Csontos, J.; Murphy, R. F.; Lovas, S. *J. Phys. Chem. B* **2008**, *112*, 3503–3508, and reference therein.

(21) Borics, A.; Murphy, R.; Lovas, S. *Biopolymers (Biospectroscopy)* **2003**, *72*, 21–24.

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

(23) Morita, S.; Fujii, A.; Mikami, N.; Tsuzuki, S. *J. Phys. Chem. A* **2006**, *110*, 10583–10590.

Quantum Mechanical Quantification of Peptide Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2885**

(24) Seifert, G.; Porezag, D.; Frauenheim, Th. *Int. J. Quantum Chem.* **1996**, *58*, 185–192.

(25) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B: Condens. Matter* **1998**, *58*, 7260–7268.

(26) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

(27) Elstner, M.; Frauenheim, T.; Suhai, S. *J. Mol. Struct.: THEOCHEM.* **2003**, *632*, 29–41.

(28) Ordejón, P.; Artacho, E.; Soler, J. M. *Phys. Rev. B: Condens. Matter* **1996**, *53*, 10441–10444.

(29) Sánchez-Portal, D.; Artacho, E.; Soler, J. M. *Int. J. Quantum Chem.* **1997**, *65*, 453–461.

(30) Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980**, *45*, 566–569.

(31) Troullier, N.; Martins, J. L. *Phys. Rev. B: Condens. Matter* **1991**, *43*, 1993–2006.

(32) Fan, W. J.; Zhang, R. Q. *Sci. China, Ser. B* **2008**, *51*, 1203–1210.

(33) Tournus, F.; Latil, S.; Heggie, M. I.; Charlier, J.-C. *Phys. Rev. B: Condens. Matter* **2005**, *72*, 075431.

(34) Tournus, F.; Charlier, J. C. *Phys. Rev. B: Condens. Matter* **2005**, *71*, 165421.

(35) Enyashin, A. N.; Gemming, S.; Seifert, G. *Nanotechnology* **2007**, *18*, 245702.

(36) Lin, C. S.; Zhang, R. Q.; Niehaus, T. A.; Frauenheim, T. *J. Phys. Chem. C* **2007**, *111*, 4069–4073.

(37) Wang, S.; Humphreys, E. S.; Chung, S. Y.; Delduco, D. F.; Lustig, S. R.; Wang, H.; Parker, K. N.; Rizzo, N. W.; Subramoney, S.; Chiang, Y. M.; Jagota, A. *Nat. Mater.* **2003**, *2*, 196–200.

(38) Cheng, Y.; Liu, G. R.; Li, Z. R.; Lu, C. *Physica A* **2006**, *367*, 293–304.

(39) Tomásio, S. D.; Walsh, T. R. *Mol. Phys.* **2007**, *105*, 221–229.

CT9002493

JCTC Journal of Chemical Theory and Computation

Binding of an RNA pol II Ligand to the WW Domain of Pin1 Using Molecular Dynamics Docking Simulations

Chai Ann Ng,[†] Daniel P. Oehme,[†] Yusuke Kato,[‡] Masaru Tanokura,[‡] and
Robert T. C. Brownlee*,[†]

*Department of Chemistry, La Trobe University, VIC 3086, Australia, and Department
of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences,
University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan*

Abstract: A novel docking protocol using a long, all atom molecular dynamics (MD) simulation, in an explicit solvent medium, without using any distance constraints is presented. This MD docking protocol is able to dock ligands, based on the C-terminal domain (CTD) of RNA polymerase II, into the tryptophan-tryptophan (WW) domain of Pin1. In this docking process, a significant loop-bending event occurs in order to encircle the ligand into its solvent exposed binding site, which cannot be simulated using current protocols. The simulations were validated structurally and energetically against an X-ray structure to confirm correct sampling of conformational space. Based on these simulations, and justification of the starting structure as a valid intermediate structure, a potential molecular basis for binding was predicted as well as confirming the key residues involved in the formation of the final strong and stable Pin1 WW domain-ligand complex.

Introduction

Human Pin1 (Pin1), a peptidyl prolyl cis/trans isomerase, was first isolated from a yeast two-hybrid screen of inter-actors of the known mitotic regulator NIMA from *Aspergillus nidulans* and shown to be essential for normal growth in some organisms.[1,2] Pin1 has been found to be involved in many cell processes such as mitosis, transcription, and in response to DNA damage. It has been seen that Pin1 is overexpressed in cancers such as breast cancer,[3] and conversely it is down-regulated in degenerative neurons in patients who have Alzheimer's.[4] Because of its involvement in these major medical issues, Pin1 has been identified as a major target for drug discovery.[5] However, Pin1 is an extremely flexible protein,[6] and this challenges the standard drug design protocol, based on the notion that binding occurs in a lock and key fashion.

Pin1 and its homologues such as PinA[7] and Ess1[8] are composed of two domains: a C-terminal peptidyl-prolyl cis/trans isomerase (PPIase) domain and an N-terminal WW

domain, which are linked by a long linker region.[9,10] It is currently hypothesized that Pin1 functions through the WW domain first acting as a targeting domain and binding the substrate before the PPIase domain isomerizes the substrate at specific phosphorylated serine/threonine-proline (pSer/Thr-Pro) motifs, either with the substrate still bound to the WW domain or, possibly, after the substrate has completely dissociated from the WW domain.[11] Named due to a pair of highly conserved tryptophan residues (at position 11 and 34 in Pin1), WW domains normally consist of 38−40 residues and fold into triple-stranded antiparallel beta sheets.[12] WW domains bind proline rich proteins and are classified into different groups depending on which motif they bind. Group IV WW domains, to which Pin1 belongs, bind to Pro residues preceded by a pSer/Thr.[13,14] Pin1 and its homologues are unique as they are the only parvulin-like proteins (a family of prolyl isomerase) with WW domains.[15] Assays on the WW domain and PPIase domain individually show that the WW domain interacts with pSer/Thr-Pro motif containing substrates with a higher affinity than the PPIase domain and of nearly equal affinity to that of the whole protein.[10,14] However in isolation, the WW domain shows no isomerase activity. It is therefore critical that the WW domain interacts

* Corresponding author e-mail: R.Brownlee@latrobe.edu.au.
† La Trobe University.
‡ University of Tokyo.

Molecular Dynamics Docking Simulations

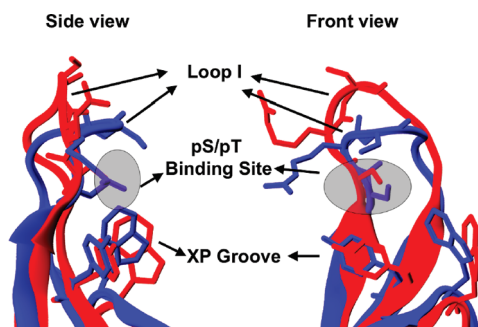*J. Chem. Theory Comput., Vol. 5, No. 10, 2009* **2887**

Figure 1. Front and side view of the superimposition of the X-ray structures of ligand free (red) and bound (blue) Pin1 WW domain. In the ligand free structure, termed the "open" form, loop I points upward. In the ligand bound structure, termed the "closed" form, the loop I bends down to cover the binding site.

with these pSer/Thr-Pro motifs containing peptides prior to the substrate being passed onto the PPIase domain for isomerization.[16]

A number of three-dimensional structures of Pin1 and its homologues are available through the Protein Data Bank[5,7−10,17,18] of which, two are extensively used in this work. The first is the X-ray structure of Pin1 bound to an AlaPro dipeptide at the PPIase domain but without a ligand bound to the WW domain (1PIN).[9] The second X-ray structure has a ligand comprising of a 7 amino acid sequence of the C-Terminal Domain (CTD) of RNA polymerase II bound to the Pin1 WW domain (1F8A).[10] This CTD ligand, (CTD-S2/S5) with sequence YpSPTpSPS, has three serines, two of which are phosphorylated, one at position 2 and one at position 5. It is of interest to note that all peptides bound to the WW domain bind with the proline in a trans conformation and that they all bind in the same unique binding mode.[11,17]

From the X-ray structures of Pin1, two important subdomains of the Group-IV WW domain were identified for ligand-binding (Figure 1). First the XP groove,[19] which is formed by residues Tyr23 and Trp34 which interacts with the proline residue of the pS/pT-P motif from the substrates in a hydrophobic fashion. Second loop I, the loop between the first and second β-sheets and composed of residues 16−21, is very flexible and binds the phosphate group immediately upstream of the proline. The flexibility of loop I is highlighted by comparing the ligand free (1PIN) and ligand bound (1F8A) X-ray structures of the Pin1 WW domain where a major difference is seen with loop I closing over the binding site in the ligand bound form (Figure 1).

A key aspect of the work in this paper is the ability to dock a ligand into a binding site. A number of programs have been developed with the premise of being able to dock a ligand into a receptors binding site. The algorithms used in these programs can be grouped according to the level of flexibility of the receptor and ligand that the algorithm allows.[20] In order of increasing complexity, the three groups are as follows: (a) rigid body docking, where both the receptor and small molecule are treated as rigid; (b) flexible ligand docking, where the receptor is held rigid, but the ligand is treated as flexible; and (c) flexible docking, where

both receptor and ligand flexibility is considered. Thus far, the most commonly used docking algorithms use the rigid receptor/flexible ligand model where the receptor site is unable to change its conformation upon ligand binding. This is mainly because treatment of backbone flexibility in protein−protein docking is quite challenging, and thus few docking programs deal with the backbone flexibility.[21] Those programs that can deal with ligand and receptor flexibility however are normally only used for screening purposes.[20] Thus they are not suitable to use in studies such as these, where a ligand is to be docked into an open form of a protein.

In this work, we were guided by Hornak et al. who used all atom molecular dynamics (MD) simulations on a HIV protease complex.[22] These long simulations, produced with an implicit solvent model and no distance constraints, were able to sample the conformational changes that take place when a ligand is placed in the active site of HIV protease, and the protease moves from an open to a closed conformation. In contrast to the HIV protease simulations, an explicit solvent model was used in this study since it is known that water molecules play an important role in the stability of the ligand in the solvent exposed binding site.[10]

Our aim in this paper was to develop a molecular dynamics protocol to simulate the large conformational change of loop I in Pin1 moving from an open form, to a closed form when a pS/pT-P motif containing ligand (CTD-S2/S5) is bound. The protocol required validation which was achieved by producing multiple simulations. The first two were duplicate simulations of the CTD-S2/S5 ligand docked into Pin1. The third, also using the MD docking protocol, was a CTD-S5 simulation where the open form of Pin1 is docked with the singularly phosphorylated CTD-S5 ligand. Finally a standard MD simulation was run on the X-ray complex of Pin1 in its native bound conformation with the CTD-S2/S5 ligand to provide standard reference structures to compare with all three simulations run with the MD docking protocol. This reference simulation allowed us to validate the protocol, confirming that the MD docking protocol was able to produce complexes that sampled correct conformational space.

The second aim of this work was to use the simulations from the MD docking protocol to propose a potential binding mechanism for the CTD ligand to the Pin1 WW domain. By justifying that the starting structure used is a valid intermediate structure, key stages in the binding process were identified in the multiple simulations and the role of key residues in the binding process and stabilization of the complex, as identified by previous biological experimental studies, were explained. Most of the investigation of Pin1 in the literature has focused on its cellular role, yet there has been no investigation into the detailed ligand binding process. The importance of Pin1 in its many cellular processes means that the results from this study will be important in the future production of ligands to inhibit Pin1.

Materials and Method

Structure Creation. The Pin1 WW domain-CTD-S2/S5 complex used as the starting structure for the CTD-S2/S5 simulations was created using the scheme seen in Figure
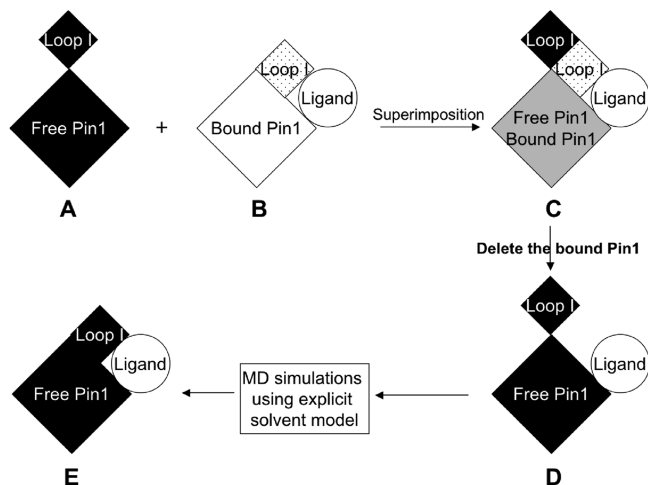
Figure 2. MD docking protocol. A→D show the production of the initial starting structure for the MD docking simulation. D→E show the initial structure moving into the closed form as seen in the "bound" X-ray structure as a result of the MD simulation.

2a−d. The "bound" structure of the Pin1 WW domain (b) (PDB ID: 1F8A), with the CTD-S2/S5 ligand (YpSPTpSPS), was superimposed onto the "free" structure of Pin1 (a) (residues 6−37 in PDB ID: 1PIN) using the Maestro program from Schrodinger.[23] From this superimposed structure (c), the "bound" Pin1 structure was deleted, and then the "free" Pin1 and the CTD ligand were merged together (d).

For the starting structure of the CTD-S5 simulation, where S2 is unphosphorylated, the Pin1-CTD-S2/S5 complex used as the starting structure for the CTD-S2-S5 simulations was modified by removing the phosphate group on S2 of the CTD-S2/S5 ligand to convert the phosphorylated serine to a standard serine residue. The crystal structure of the "bound" form of Pin1 (1F8A) was used as the starting structure for the X-ray complex simulation.

Parameters for the standard amino acids in the Pin1 WW domain and the ligand were taken from the AMBER ff03 force field,[24] while parameters for the phosphorylated serines (phosphoserine with deprotonated phosphate group, SER-PO3) were taken from the AMBER Parameter Database (http://pharmacy.man.ac.uk/amber/).[25] Hydrogen atoms were added such that all residues were in their standard protonation states at pH 7. Each complex was placed in a octahedral box of TIP3P water molecules[26] such that the box extended at least 10 Å from any atom of the Pin1-CTD ligand complexes. The systems were neutralized and made up to a salt concentration of 0.2 M by adding appropriate numbers of Na$^+$/Cl$^-$ ions. These final steps of molecular system creation were achieved using the Leap module of AMBER9.[27]

MD Simulation Protocol. All simulations involved in the MD docking protocol were performed using the AMBER9 package of programs.[27] A four-step equilibration scheme was performed with the first step being a minimization where the Pin1/ligand complex was held fixed with a restraint of 500 kcal mol^{-1} Å^{-1}, such that the position of water molecules and ions could be relaxed. A second minimization followed, whereby all atoms were allowed to move for 1000 steps of

steepest descent and then 1000 steps of conjugate gradient to relax the system as a whole. Once sufficiently relaxed, the system was carefully heated to 300 K over 10 ps using a Langevin dynamics temperature regulation scheme[28] with a collision frequency of 1 ps^{-1}, with everything except water held under a 10 kcal mol^{-1} Å^{-1} restraint. Volume was kept constant, and SHAKE[29] was used to constrain bonds involving hydrogen such that a time step of 2 fs could be used. The final step of the equilibration scheme, to equilibrate the physical parameters such as temperature, pressure, and density of the system, was run over 10 ps using NPT conditions. Temperature and pressure were held constant using Langevin and isotropic scaling schemes, respectively. Production phase was run in 10 ns blocks under the same NPT conditions used for the final step of equilibration. The random seeds used to generate velocities were changed for each different block of a simulation to stop the possibility of synchronization of trajectories.[30] SHAKE was used, and long-range electrostatic interactions were calculated using the PME method with a cutoff of 10 Å.[31] Coordinates were printed to the trajectory file every 500 steps (every 1 ps) such that each trajectory held 10000 snapshots and a restart file was written every 1000 steps. Each simulation was continued on for a further 10 ns after a stable complex was formed for validation purposes.

Structural Analysis. Visual, distance, torsion, and rmsd analyses were completed using VMD.[32] All quantitative data used in the validation of the MD docking simulations were taken from the last 10 ns of the simulations. Over this time it was confirmed that the simulations were sampling a stable conformational space through hydrogen bond (H-bond) analysis using the ptraj module of AMBER9 (data not included).

Distances between the S5 phosphate of the CTD ligand, and key functional groups in loop I which are shown to H-bond to S5, were calculated between the phosphorus of the S5 and the hydrogen connected to the H-bond donor in loop I. Distance data for the Pin1 1F8A structure were calculated using the final structure from the first minimization of the X-ray complex from the X-ray complex simulation. This minimized structure was used instead of the initial X-ray structure as the X-ray structure did not include hydrogen atoms in optimized positions. The phi/psi angles on either side of the amide bond between Arg17 and Ser18 were calculated to quantify amide bond flipping.

The X-ray structure of the Pin1 complex (PDB ID: 1F8A) was used as the template structure for all rmsd calculations. All trajectories were aligned onto the backbone of this structure, and rmsds were calculated for each snapshot over the entire production phase. Snapshot intervals were changed to 5 ps to ensure that snapshots would be uncorrelated when compared to the next/previous snapshot.[33] Heavy atom and backbone rmsds were calculated for the entire Pin1 WW domain (residues 6−37), the loop I residues (16−21), and the XP groove (23 and 34) and were then averaged for each 10 ns portion of production phase.

Root mean square fluctuations (RMSF) for each residue were calculated using ptraj (see the Supporting Information). Each 10 ns portion of a simulation was rms fitted, for residues
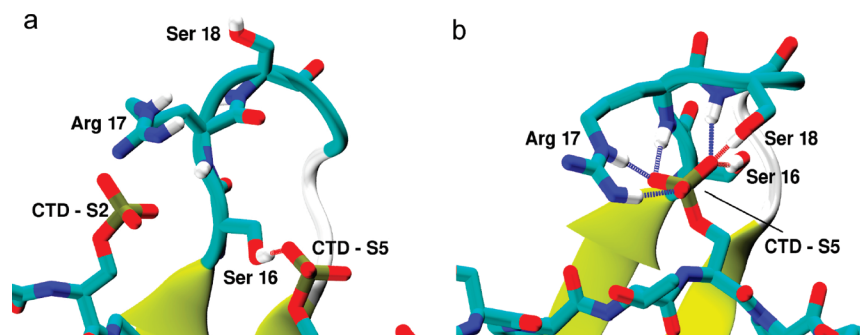
Figure 3. Interactions between loop I of the Pin1 WW domain and the CTD S2/S5 ligand in (a) the open and (b) closed form of the Pin1 complex from the MD docking protocol simulation. Specific hydrogens are shown to highlight the 6 key hydrogen bonds shown in (b) that stabilize the bound form of the complex.

6−37 from Pin1, onto the first structure of that portion of the simulation. RMSF were calculated for both Pin1 and CTD ligand residues.

Binding Energy Calculations. Binding energies were calculated using the MM-PBSA approach from AMBER9.[34,35] The single trajectory approach for calculating MM-PBSA energies was used with the ensemble of structures obtained from the final 10 ns of each simulation. Snapshots were taken every 5 ps from these trajectories such that 2000 snapshots were used in the calculations. For each snapshot, water molecules and counterions were removed, and estimates of the binding energies were calculated using

$$\Delta G_{bind} = G_{complex} - (G_{receptor} + G_{ligand}) \qquad (1)$$

where the free energy of each part of the system was calculated using the formula

$$G = <E_{MM}> + <G_{solv}> - <TS_{solute}> \qquad (2)$$

with <> signifying average values over the simulation. The E_{MM} term incorporates the molecular mechanical (MM) energies of the system. Using the single trajectory approach, this term was calculated using just the nonbonded, electrostatic and van der Waals components. The solvation energy (G_{solv}) term was divided into two parts

$$G_{solv} = G_{pol} + G_{np} \qquad (3)$$

The nonpolar contribution (G_{np}) to G_{solv} was calculated using

$$G_{np} = \gamma SASA + \beta \qquad (4)$$

where $\gamma = 0.0072$ kcal mol^{-1} Å^{-2}; $\beta = 0.0$ kcal/mol; and SASA is the solvent accessible surface area. The polar contribution (G_{pol}) to G_{solv} was calculated using both the Poisson−Boltzmann (PB) and Generalized-Born (GB) methods. PB estimates of G_{solv} were calculated by solving the Poisson−Boltzmann equation using the PBSA module of AMBER 9. GB estimates of G_{solv} were calculated using the GB model developed by Onufriev et al. and implemented in SANDER from AMBER 9.[36] The interior and exterior dielectric constants were set to 1 and 80, respectively, with a grid spacing set to 0.5 Å and 1000 linear iterations were performed. Bondi radii and a probe radius of 1.4 Å were used for both G_{pol} and G_{np} calculations. The entropy

component (TS_{solute}) of the system was not calculated as discussed in the Results section of this paper.

Results and Discussion

The MD docking protocol was tested by running three separate simulations. The first test involved running on a simulation on the Pin1 CTD-S2/S5 complex. Further testing of the MD docking protocol was achieved by performing an independent duplicate of the CTD-S2/S5 simulation. The independence of the duplicate simulation was achieved by changing the random number seeds which initiate velocities and the forces on atoms at each stage of the simulation. This guaranteed that although the simulations started from the same structure, the velocities generated would be independent and therefore the simulations also. It has also been shown that if the random number seed values are not changed, simulations can synchronize.[30]

Binding energy studies have shown that there is no significant difference in binding affinity between Pin1 binding to the CTD-S2/S5 and CTD-S5 ligands,[10] suggesting that phosphorylation of S2 has a limited role in the binding process and therefore the bound structure would resemble the X-ray structure of the CTD-S2/S5 complex. The third and final test of the MD docking protocol was to run a simulation on the Pin1 CTD-S5 complex to determine whether the protocol was able to predict a bound conformation of this new complex, and, if so, did it have a similar conformation to the bound CTD-S2/S5 ligand complex.

In all three MD docking simulations performed, Pin1 was seen to fold from its open form (Figure 3a), into an intermediate "semi-open" form of the Pin1 WW domain after an 180° amide flip between Arg17 and Ser18 in loop I, to finally the closed form (Figure 3b), creating a ligand binding pocket as seen in the X-ray crystal structure 1F8A.

Validation of the MD Docking Simulations. The quality of the simulations, and in particular, the final structures in which the ligands were bound to Pin1, were validated through comparisons with the X-ray structure of the bound complex (1F8A). It should be noted that this X-ray structure was not used to generate the simulation. rmsd, distance, and binding energy comparisons were made from the structures from the final 10 ns of each simulation (2000 structures) to the X-ray structure.

The data from Table 1 show that for the backbone atoms of the whole protein, loop I, and the XP groove, the rmsds

Table 1. Rmsds over the Last 10 ns of Each Simulation against the X-ray Structure of the Pin1 Complex (1F8A)[a]

	all		loop I		XP groove	
	backbone	heavy	backbone	heavy	backbone	heavy
CTD-S2/S5	0.97 (0.07)	1.97 (0.09)	1.01 (0.16)	1.53 (0.11)	0.77 (0.09)	1.20 (0.03)
CTD-S2/S5 (duplicate)	1.02 (0.08)	2.17 (0.08)	0.87 (0.10)	1.61 (0.17)	0.72 (0.08)	1.23 (0.05)
CTD-S5	1.04 (0.11)	2.00 (0.12)	0.99 (0.13)	1.37 (0.19)	0.69 (0.06)	1.18 (0.02)
X-ray complex	0.78 (0.11)	2.13 (0.25)	1.10 (0.31)	1.43 (0.20)	0.62 (0.12)	0.67 (0.08)

[a] Standard deviation in parentheses.

Table 2. Distances (Å) from the Phosphorus of S5 to the Hydrogens of Loop I Functional Groups over the Last 10 ns[c]

	Ser16 OH	Arg17 N	Arg17 NE	Arg17 NH	Ser18 N	Ser18 OH
Pin1 1F8A[a]	2.93	2.98	2.88	2.47	3.32	5.57
CTD-S2/S5	2.85 (0.03)	2.74 (0.06)	2.85 (0.03)	2.66 (0.06)	3.27 (0.26)	3.17 (0.75)
CTD-S2/S5 (duplicate)	2.82 (0.03)	2.72 (0.05)	2.86 (0.05)	2.72 (0.08)	3.10 (0.05)	2.81 (0.18)
CTD-S5	2.84 (0.03)	2.69 (0.07)	2.79 (0.04)	2.70 (0.09)	3.03 (0.13)	2.90 (0.16)
X-ray complex	2.81 (0.03)	2.67 (0.01)	2.78 (0.01)	2.65 (0.01)	3.13 (0.02)	2.81 (0.02)
initial[b]	2.16	5.78	10.41	11.83	9.00	10.51

[a] Pin1 1F8A distances are taken from the final structure of the first stage of minimization from the X-ray complex simulation. [b] Initial distances are taken from the structure used to start all the CTD simulations. [c] Standard deviation in parentheses.

Table 3. Binding Energies from Last 10 ns of Simulations[a]

method	CTD-S2/S5	CTD-S2/S5 (duplicate)	CTD-S5	X-ray complex
PB	−64.18 (1.05)	−64.05 (0.92)	−63.81 (0.89)	−66.23 (0.99)
GB	−52.42 (0.80)	−53.57 (0.75)	−53.65 (0.73)	−55.92 (0.78)

[a] Energies in kcal mol⁻¹ and standard errors in parentheses.

are all about 1 Å. This is well below the standard 2 Å limit with which structures are said to be "nativelike".[37] In order to ascertain the quality of this data, a MD simulation of the X-ray structure (1F8A) with hydrogens added and in a solvent box was run and the last 10 ns of the simulation compared back to the single X-ray structure (shown in bottom row of Table 1). These data are almost identical to the data seen for the simulations run with the MD docking protocol suggesting the simulations are sampling a conformational space similar to the X-ray structure.

Interactions between the loop I residues in Pin1 and the S5 phosphate of the CTD ligand are compared by calculating the distances of six important hydrogen bonds from the last 10 ns of the simulations against the distances seen in the X-ray complex. Table 2 shows that five of the six hydrogen bonds between the loop and the ligand are directly comparable with the X-ray structure. For Ser18 OH however, there is a considerable difference between the distance in the simulation (3.17 Å) and the distance in the X-ray structure (5.6 Å). Interestingly, the X-ray complex simulation shows this considerable difference to the single X-ray complex structure as well and is discussed below. For completeness, Table 2 also shows the six H-bond distances from the initial structures used for the CTD-S2/S5 simulations where loop I was in its open form. These highlight the dramatic changes in structure that took place in the simulation. Overall the six H-bond distances are internally self-consistent and comparable with the X-ray structure confirming that our simulations were sampling the correct conformational space.

The final tool used to validate the MD docking simulations was the calculation of "binding energies" using the MM-PBSA protocol within the AMBER software.[34,35] The binding energies of the MD docking simulations and the X-ray complex simulation are very similar, differing by on average 2.2 and 2.7 kcal/mol for the Poisson−Boltzmann (PB) and Generalized Born (GB) methods, respectively (Table 3). The results from the PB and GB methods are expected to be different but show internal self-consistency[38] which is what is seen in these results. It is important to note that these energies are purely relative and are not absolute estimates of the binding energy. MMPBSA/GBSA has consistently been shown to be a good method for comparing binding energies of similar complexes.[34,38,39] Entropy calculations have not been included as it has been documented in many papers that it is difficult to calculate, and with similar ligands it can be assumed that differences in entropy will be small.[34,40] Given that the simulations started from structures very different to the X-ray structure, and that such close agreement is seen, the binding energy data again confirms that the simulations are sampling a very similar conformational space.

Importance of an Explicit Solvent Model. The use of an explicit water model was essential to simulate the transformation of the Pin1 WW domain from the "open" to "closed" form. Preliminary simulations performed on the CTD-S2/S5 complex following the protocol of Hornak et al., using an implicit solvent model failed consistently with the CTD ligand quickly flying away from the binding site. These simulations failed due to the binding site being on the surface of the protein and, hence, solvent exposed, which contrasts to the HIV protease work where the ligand is placed in a cavity which is solvent excluded. By modeling the Pin1/CTD ligand simulations in explicit solvent, the movement of the ligand was significantly restrained in the binding site which allowed loop bending and, eventually, the formation of the strong and stable complex.

The crystal structure of the Pin1 complex (1F8A) showed that there is a key water mediated interaction between Tyr23 and the S5 phosphate of the ligand. Due to the solvent exposure of the binding site, it is expected that many different water molecules will mediate this interaction. This is seen in the MD docking simulations with the water molecules continually exchanging with the bulk solvent. This continual

Molecular Dynamics Docking Simulations

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2891**

exchange of water molecules allows for a consistent interaction between Tyr23 and S5, helping to stabilize the complex at early stages and increase strength of binding during the latter stages. This again illustrates the importance of performing these simulations in explicit solvent as an artificial restraint would need to be placed on the structural water molecule for it to stay in position in an implicit solvent simulation.

Successful simulation using the MD docking protocol relied upon placement of proline 6 of the ligand in the XP groove of Pin1, with the proline sandwiched between Try23 and Trp34. It could be suggested that this is a limitation of the MD docking protocol in that we have used the bound X-ray structure as a basis for this work. However we view the work in this article is a proof of concept showing that we can position a ligand with one key component near the correct position and run a simulation to produce a known bound complex.

Overall it has been shown that the MD docking protocol has been able to dock ligands into a binding site and simulate the movement of the complex from an "open" free state to a "closed" bound state. It has also predicted that the CTD-S2 ligand, of which no crystal structure has been produced, binds in an identical manner to the CTD-S2/S5 ligand, as highlighted by similar rmsd, distances and pleasingly having a similar binding energy as seen previously in experimental studies.

Proposed Binding Mechanism

Justification of Initial Ligand Placement. If the starting structure used in the MD docking simulations is a valid intermediate in the binding process, mechanistic information from the MD simulations can provide a detailed description of the binding process. The simulations presented in this paper start with the ligand proline (Pro6) positioned close to the XP groove. Given that the predominant ligand interactions with Pin 1 in the bound structure are (i) the Pro6/XP groove interactions and (ii) the phosphorylated serine (pSer) interactions with the flexible loop I, it might be argued however that an alternative starting geometry based on the pSer/loop I interaction could be plausible.

There is significant evidence supporting the starting geometry used in these simulations as a valid intermediate. The first reason is by docking Pro6 into the XP groove, significant hydrophobic burial occurs, in which the placement of the hydrophobic proline in XP groove reduces the unfavorable solvent contacts these hydrophobic residues experience. Second, by docking Pro6 into the XP groove, the pSer is positioned in the vicinity of the loop binding site as a result of the geometric consequences of Pin1 only binding the trans form of the CTD ligand.[11,17] Placing pSer near the binding site allows subsequent bending of loop I to generate a number of H-bonds to pSer. Third, mutational analysis of the XP groove residues shows no binding is seen when either residue, Tyr23 or Trp34, is mutated to alanine.[14] This suggests that binding can only occur when a fully functional XP groove is available. Finally the fact that our MD docking simulations have been able to correctly produce
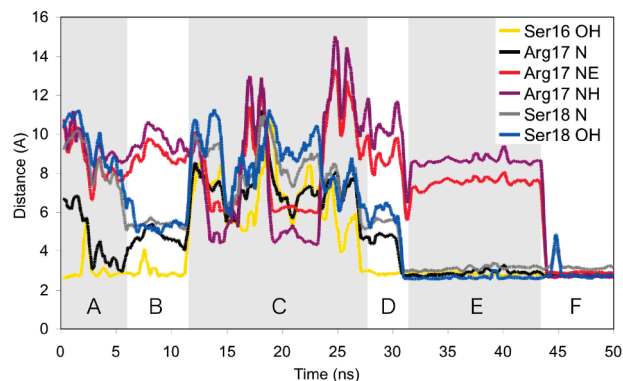

Figure 4. Distances from key functional groups in loop I residues to the phosphorus atom of S5 of the CTD ligand for the first 50 ns. The timeline is shaded to define the six stages of the simulation. Distances were calculated as running averages over 100 snapshots. After 50 ns, only insignificant changes are seen and have therefore been excluded from the figure for clarity.

structures that sample a similar conformational space to the X-ray structure suggest this could be a valid intermediate structure.

The alternative starting geometry which relies upon pSer first interacting with loop I of Pin1 is not favored as a valid intermediate for a number of reasons. The first is that pSer is a polar residue which would result in a significant desolvation penalty which pSer would have to overcome upon binding by forming interactions with a number of residues simultaneously.[41−44] This is unlikely since loop I is not optimized to interact in its open form with pSer as: Ser16 is hidden in the binding site and is not accessible; Ser18 points out toward the solvent; while Arg17 is flapping around in space, far from the binding site. Another reason is that any interactions between pSer and loop I would significantly decrease the flexibility of loop I,[45−47] resulting in significant barriers to allow loop I to bend, and to bring the CTD ligand along with it. Finally, mutational analysis of key loop I residues to alanine show that although a reduction of binding strength is seen, binding occurs none the less.[10,14,48] Therefore by removing potential H-bond partners, binding still occurs suggesting initial binding may not take place via this approach.

CTD-S2/S5 Docking Simulation. Given that the starting structure has been shown to be a valid intermediate, the MD docking simulations could be investigated to identify a potential binding mechanism. The final structures from the initial CTD-S2/S5 MD docking simulation show that there were 6 key hydrogen bonds (H-bonds) between residues in loop I and the CTD ligand (as shown in Figure 3b). The progress of binding was followed by monitoring these distances throughout the simulation, in which 6 distinct stages are clearly identified (Figure 4). This MD docking simulation is much more readily understood by visualization and therefore a QuickTime movie (15MB) of this simulation is available in the Supporting Information for this paper. Each stage is characterized by a geometric change in the structure of the complex. The stages, A through F, of the simulation are described below and give a fascinating insight into the proposed binding mechanism.
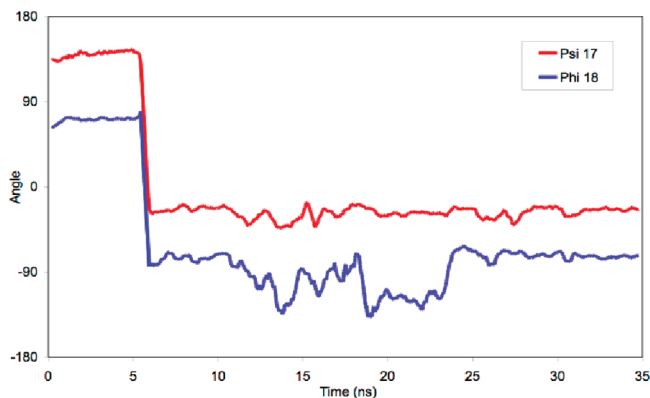
Figure 5. Plot of the psi 17 (17N - 17C-alpha - 17C - 18N) and phi (17C - 18N - 18C-alpha - 18C) torsion angles from the first 35 ns of the simulation. Angles were calculated as running averages over 100 snapshots.

A) Starting Conformation. At the start of the simulation the proline at position 6 of the CTD-S2/S5 ligand is positioned into the XP groove of Pin1 according to the bound X-ray structure. Throughout the whole simulation there is very little movement in this proline, with this group functioning as an anchor, holding the ligand in place while other motions takes place. With the proline sandwiched in between Tyr23 and Trp34, the S5 phosphate points up toward loop I and forms a stable hydrogen bond with the Ser16 OH. Loop I is open at this stage of the simulation, and no other hydrogen bonds are seen between the S5 phosphate and the rest of loop I.

B) Amide Flip. After about 6 ns the amide bond between Arg17/Ser18 flips, as shown by the 180° change in the psi-17/phi-18 angles in Figure 5. This flip of the amide bond does not significantly alter the backbone of adjacent residues nor does it change the omega angle of the amide bond (data not shown). The carbonyl of Arg17 now points out toward the solvent while the amine (N−H) of Ser18 points toward the ligand, and loop I bends into a "semiopen" conformation. Thus the distances between both the backbone Ser18 NH (Ser18 N) and the Ser18 OH and the S5 phosphorus decrease from about 8 Å to just above 5 Å (Figure 4). Following the flip, a short period of stability (from 8−12 ns) is seen in which a number of water molecules position themselves between loop I and the S5 phosphate, forming water mediated H-bonds. These water molecules effectively block S5 from interacting directly with the loop I residues, in turn blocking the transition to the "closed" form. As a result of this, instead of moving closer toward loop I, at about 12 ns the S5 phosphate moves away from the binding site and a period of chaos ensues.

C) Chaos. During this period, no hydrogen bonds are seen between loop I and the S5 phosphate, and a high degree of flexibility is seen throughout the complex. From RMSF data (Supporting Information) it can be seen that the backbone of S5 is quite stable with the side chain being flexible. The flexibility of the S5 side chain is highlighted in this "chaos" period where it rotates about 60° out into the solvent, as shown in Figure 6, allowing layers of water molecules to position themselves between S5 and loop I. During this chaos period water mediated interactions are seen between S2 and
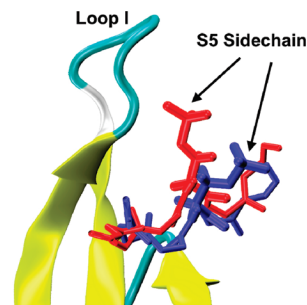

Figure 6. Pin1 WW domain in its semiopen form with CTD-S2/S5 ligand having its S5 side chain positioned in the pS/pT binding site (red) and after 60° rotation away from the binding site in the "chaos" period (blue).

S5, interactions between S2 and the side chain of Arg17, and interactions between S5 and the side chain of Arg17. After 16 ns of chaos, S5 rotates back into the binding site, as shown by the decreases in length between Ser16 OH and S5, and chaos finally comes to an end when a stable H-bond again forms between these two residues at about 28 ns.

D) Bending Cascade. The reformation of the S5 to Ser16 OH H-bond starts a cascade of events which allow the loop to fully bend over into the "closed" conformation. The first step of this cascade occurs at 30 ns where the water mediating the interaction between the Arg17 backbone NH (Arg17 N) and the S5 phosphate moves from the binding site allowing a direct H-bond. This causes a small downward movement of loop I which in turn removes the waters mediating interactions between Ser18 N/Ser18 OH and the S5 phosphate. This causes another downward movement of the loop and two direct H-bonds between Ser18 and S5 are formed. Thus at 32 ns we have 4 direct H-bonds to S5 phosphate with all distances being consistently 3 Å, and therefore the complex is in a stable conformation.

E) Arg17 Side Chain Optimization. During the time that the rest of loop I was stable, the side chain of Arg17 alternately interacts with both the solvent and with the S2 phosphate. At ∼44 ns, the side chain of Arg17 finally forms two H-bonds with the S5 phosphate. After this time in the simulation, the Arg17 side chain remains at the standard 3 Å distance from the S5 phosphate.

F) Strong Stable Complex. Throughout the remaining 25 ns (through to 70 ns) a predominantly strong and stable complex is seen. Occasionally loop I tries to move back to the semiopen form, as seen by the Ser18 distance increasing to ∼5 Å. This is quickly rectified, and the Ser18 distances decrease back to 3 Å almost immediately, presumably because of the stability caused from the rest of the H-bonding interactions, such that the strong stable complex is the predominant conformation seen in the last 25 ns of the simulation.

CTD-S2/S5 Duplicate Docking Simulation. Although the same conformational space is being sampled at the end of the duplicate simulation, Figure 7 shows there are considerable differences in the time evolution of the two simulations, in particular that there is no "chaos" period (stage 'C') in the duplicate simulation. This is because S5 does not move outside the binding site and is shown in Figure 7 by the Ser16

Molecular Dynamics Docking Simulations

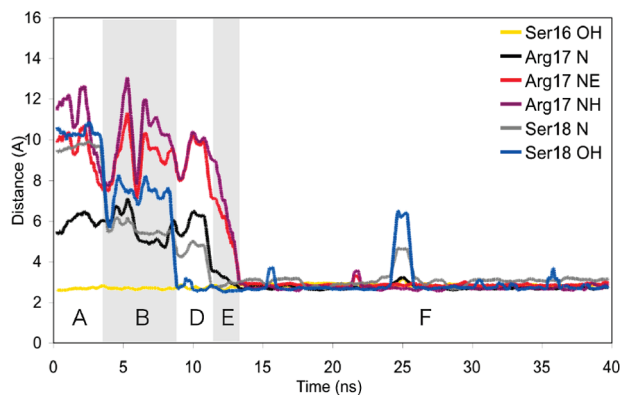*J. Chem. Theory Comput., Vol. 5, No. 10, 2009* **2893**

Figure 7. Distances from key functional groups in loop I residues to the phosphorus atom of S5 of the CTD ligand for the duplicate simulation. The timeline is shaded to define five out of the six stages of the simulation as defined for the initial CTD-S2/S5 simulation. Distances were calculated as running averages over 100 snapshots.

Figure 8. Distances from key functional groups in loop I residues to the phosphorus atom of S5 of the CTD ligand for the CTD-S5 simulation. The timeline is shaded to define the six stages of the simulation. Distances were calculated as running averages over 100 snapshots.

OH distance never deviating from around the 2.7 Å distance. After the amide flip, which occurs on a similar time scale to the initial simulation, S5 stays in the binding site, and the bending cascade occurs much earlier. In the duplicate we also see that side chain optimization occurs at a much earlier stage than the initial simulation. Overall this means that a stable complex is formed after only ~15 ns compared to the 45 ns it took in the initial simulation. Although no "chaos" period is seen in the duplicate simulation, all other five stages of the binding can be identified.

An interesting aspect of the duplicate simulation is that the bending cascade occurs in the opposite order to the initial simulation. Instead of interacting with Arg17 N first and then interacting with Ser18 N and Ser18 OH, S5 H-bonds to Ser18 OH first and then to Ser18 N followed by Arg17 N. Visual analysis shows that this is because for the first part of the simulation, S5 interacts with Arg21 of loop I. This causes S5 to sit in the far right-hand side of the binding site (Figure 1-front view). In this position, S5 is unable to interact with Arg17, and it is only when Ser18 OH starts to interact with S5 (~9 ns) that we see a shift in the position of S5 back to the center of the binding site. This movement allows Ser18 N to H-bond to S5 which in turn allows Arg17 N to interact. With S5 centrally located in the binding pocket, side chain optimization can then occur. It is important to see the movement of S5 from interacting with Arg21 at one end of loop I to eventually interacting with Arg17 as it shows that although there are some favorable interactions between Arg21 and S5, it is more favorable for S5 to sit in the left-hand side of the binding site where multiple interactions can occur which will allow for a more stable complex.

As this duplicate has been started with a different seed compared to the initial simulation, the velocities of each atom at the start of the simulations will be different. Because of this, the different simulation will follow different trajectories. It is therefore to be expected that the simulations follow different paths and are not identical. However, although the simulations do not follow the same path, they do both show the key steps of binding: amide flip; a loop I bend; and Arg17 side chain optimization to derive bound structures as close
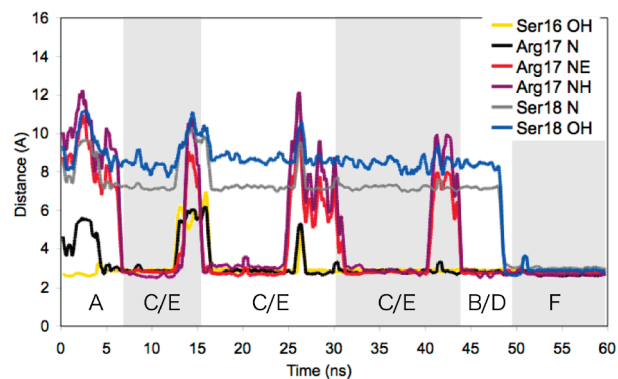
to identical as possible to the X-ray structure, thus highlighting the quality of the MD docking protocol.

CTD-S5 Docking Simulation. The MD docking simulation of this mutated ligand showed that despite the absence of phosphorylation of Ser2 of the CTD ligand, the system was able to sample a conformational space very close to that of the X-ray structure of the CTD-S2/S5 complex. The binding process differed however from that seen for the initial MD docking simulation and its duplicate, although all six stages can be seen in the trajectory (Figure 8). A striking feature of this "unphosphorylated" simulation is that the amide flipping is the last conformational change to occur and does not occur until ~48 ns which is more than three times longer than any other simulation.

As it takes so long for the amide to flip, a number of conformations of the complex are seen in the preceding time. It can be seen from the distance graph (Figure 8) that after ~5 ns, both Ser16 OH and Arg17 N are H-bonded to S5. This indicates that Arg17 N can bind to S5 at any stage of the whole binding process and is not dependent on loop bending as suggested from the initial simulations. Unexpected interactions between the Arg17 side chain and S5 (stage 'E') are also seen before loop I bending. Due to the amide flip and loop I bending having not occurred yet, these interactions are not stable, and, thus, the system is able to move into small stages of chaos. It is thought that these stages of chaos are shorter than seen in the initial simulations due to S2 no longer being phosphorylated and therefore not being able to stabilize the Arg17 side chain or S5 in their "chaotic" forms, allowing them to position themselves back to the binding site relatively quickly.

In the end it appears coincidental that at the time that the amide flip occurs, all other important residues are H-bonded to S5 which allows immediate formation of the strong and stable complex. This is different than the other simulations where the amide flip occurred before S5 was in its "bound" position. It can therefore be concluded that the bending cascade (stage 'D') can only occur once the amide has flipped and the S5 is positioned correctly in the binding site. This shows the pronounced importance of the amide flip and how it is a rate limiting step of the proposed binding mechanism.

Key Residues in Simulations. MD docking simulations in this work have provided significant insight into a potential binding mechanism of CTD ligands with the pS/pT-P motif to the Pin1 WW domain. In particular it highlights the importance of the structure of the complex around the XP groove; the amide flip; and the interactions between S5 and the loop I residues.

The MD simulations produce in this work show that the structure in the vicinity of the XP groove is very stable, as seen from the backbone data in Table 1 with an average rmsd of ~0.7 Å. This compares favorably with the rmsd from the X-ray complex simulations and given the low backbone rmsd to start with, it can be concluded that the backbone of the XP groove must have started in a conformation very similar to the final bound conformation and does not deviate significantly from this position. It is appropriate to suggest that the stability of the XP groove is due to its hydrophobicity and it being unfavorable for it to interact more significantly with the surrounding polar environment.

Other than interacting with proline, Tyr23 plays another key role in binding by being able to also form water mediated H-bonds to S5. Interaction energy calculations show that there is a significantly high electrostatic interaction between Tyr23 and the S5 phosphate. These energy calculations are unable however to include the contribution that the water mediated H-bonds play as they are transient (different water molecules play the role of the mediator). It can also be seen from these calculations that there is a significant VDW contribution to the interaction energy between proline for both Tyr23 and Trp34. No doubt this is in part due to their hydrophobic interactions to the proline from the ligand. Together with the data presented earlier, this work has shown the critical role that the XP groove has in binding and the production of a stable complex.

A significant difference in the orientation of the amide bond between Arg17 and Ser18 is observed between the open (1PIN) and closed (1F8A) forms of Pin1 WW domain. In the open form, the NH points away from where the ligand binds, while, in the closed form, the NH points toward the ligand. In all the MD docking simulations produced in this work, this amide 17/18 flip was seen allowing a conformational change of loop I to occur unhindered.

This amide bond flipping is a very important step in the binding process of the ligand to Pin1, as the loop cannot bend until this amide flip has taken place. Without the amide flip, interactions between S5 and Ser18 N or Ser18 OH would not be seen which would severely limit the ability of a strong and stable complex to be formed. To consolidate this hypothesis, a further simulation was run in which the amide bond in the free/open structure was manually flipped such that it showed no structural difference to its closed form. From this simulation, conformational change of loop I was seen after just 1 ns, which gave strong evidence to suggest that the NH of the amide between Arg17 and Ser18 must be pointing toward the ligand for conformational change to occur. Therefore the timing of the amide flip is critical to the timing of conformational change and therefore binding.

Initial comparison of the Pin1 free and bound complexes showed the greatest difference between the two structures was seen in the loop I region. This has been highlighted in our simulations and through rmsd, distance, and RMSF calculations, these changes have been quantified. All this is understandable as this is where the important loop bending occurs which locks the ligand into its bound position.

Of the loop I residues, the first residue of the loop, serine 16, is the most stable. The reduced movement seen in Ser16 relative to the other residues suggests that Ser16 may initially be in an optimal position for binding. Our simulations show that Ser16 consistently interacts with S5, and the interaction is necessary for subsequent stages of the binding process to occur. Without interaction between Ser16 and S5, no other interactions are seen between S5 and the loop except for the Arg17 side chain for small time scales in the chaos period of the initial CTD-S2/S5 simulation. It therefore seems that Ser16 is crucial to the whole binding process, which has been suggested in other work.[10,19,48,49]

Previous work, based on the crystal structure of the bound form of Pin1, suggests that Ser18 plays a minimal role in the binding to the CTD ligand.[10] In that X-ray structure (which does not include hydrogens), the Ser18 backbone NH is slightly outside the H-bonding distance, and the oxygen in the Ser18 side chain points away from the ligand, thereby preventing any H-bonding to S5. It is because of these distances, and some binding energy calculations,[48] that Ser18 was predicted to be unimportant. However Ser18 is one of only two residues from loop I that are conserved in the analogues of Pin1, PinA, and Ess1, suggesting a more significant role in the binding process. In addition to this, our MD simulation of the X-ray complex, which started from the initial X-ray structure, actually shows the Ser18 side chain rotating such that the distance from the S5 phosphorus to Ser18 OH decreases from 5.57 Å to within 3 Å. The Ser18 backbone NH also sits at a similar distance, ultimately allowing two H-bonds to form to S5. The MD docking simulations carried out also show this conformation where the Ser18 NH/OH are within H-bonding distance of S5 (Table 2). Given that this form of the complex is the predominant one seen, the simulations indicate that Ser18 has a much more significant role to play in binding than previously thought.

Since the X-ray structure of Pin1 bound to the CTD ligand was published it has been thought that Arg17 played a very important role in binding.[10] Due to arginine being a positively charged amino acid and the S5 phosphate from the ligand being negatively charged, it is easy to understand why this has been thought. This has been supported by binding studies which have shown that by mutating Arg17 to Ala, the binding strength decreases 6-fold,[10] and that a homologue of Pin1, PinA, where Arg17 is replaced by Asn, also has a binding strength ~6-fold less.[19] Our simulations also support the importance of Arg17 to the binding strength by showing that Arg17 contributes 3 H-bonds to the final complex, two from the side chains and one from the backbone NH.

Arginine is a highly charged, flexible residue that is regularly seen on the surface of proteins. The ability of our simulations to show this flexibility, yet also to show the remarkable stability when the Arg17 side chain interacts with S5 is extremely important to the successful simulation of

Molecular Dynamics Docking Simulations

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2895**

the binding process. The fact that different time intervals are seen for when this stability occurs also highlights the fact that the Arg17 side chain optimization step is an independent process and another rate-limiting process in binding of the Pin1 WW domain to a ligand.

As important as Arg17 has been shown to be to the overall binding strength of Pin1 to the CTD ligand, our simulations show that Arg17 is not critical for binding to occur. There are numerous stages in the simulations where a stable complex is formed without interaction between S5 and the Arg17 side chain and where Arg17 interactions occur yet stability is not seen. The binding studies discussed above also suggest that Arg17 is not critical as complexes are still formed with acceptable binding energies when Arg17 is not present. Sequence alignment of the three analogous WW domain containing proteins, Pin1, PinA, and Ess1, show that the key residues in loop I at positions 16, 17, and 18 are conserved with residue 17 either an arginine, asparagine, or lysine. This suggests that the binding motif that pSer looks for may be S-X-S, where X is an amino acid containing at least one amine group in its side chain.

Conclusion

The development of an MD docking protocol using all atom unrestrained MD simulations in this work has allowed us to simulate the binding of a ligand into a solvent exposed binding site. Validation of the MD docking simulations has shown that our MD docking protocol has produced structures that sample the same conformational space as the X-ray structure of the bound complex. The simulations also highlight the importance of using an explicit solvent model for docking a ligand into a solvent exposed binding site.

The results of the simulations, given that the starting structure is a valid intermediate as has been justified, show that the binding of a ligand containing the pS/pT-P motif to Pin1 is a multistage process. First, the proline of the ligand docks itself in between the XP groove residues Tyr23 and Trp34 of Pin1 where it is stabilized by hydrophobic interactions. The next steps in the binding mechanism are the flip of the amide bond between Arg17 and Ser18 and Arg17 side chain optimization. The flipping of the amide bond allows loop I to move into a semiopen conformation, from which it eventually moves into a closed conformation when S5 is correctly positioned in the binding site. Optimization of the Arg17 side chain occurs when H-bonds form between the NHs of the Arg17 side chain and the oxygens from the S5 phosphate. It is only when all stages of the binding mechanism have been completed do we see the strong and stable complex.

Our work using a MD docking protocol for solvent exposed binding sites has shown that a number of interactions between the pS-P motif in the ligand and the XP groove and loop I residues are vital for the formation and strength of the bound complex. These results confirm that for a ligand to bind into the Pin1 WW domain, it must have a hydrophobic fragment to bind into the XP groove, connected to a fragment containing a phosphate or some very electronegative atoms in the correct spatial relationship to interact with

loop I as highlighted by recently published work involving phosphorylated serine mimics.[50]

This work has shown that the MD docking protocol can be used as a refining tool after docking. To become a truly predictive tool, the protocol should be able to place the CTD ligand in any conformation in the environment surrounding the binding site and, following a MD docking simulation, have the complex settle into the bound conformation. It would then be possible to use the protocol to identify potential ligands designed to inhibit phospho-peptide binding.

Supporting Information Available: A QuickTime movie of the CTD-S2/S5 MD docking simulation; additional data from the MM-PBSA calculations; figures detailing the change of RMSF over time for residues in Pin1 and in the CTD-ligand; and the complete ref 27. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Joseph, J. D.; Daigle, S. N.; Means, A. R. PINA is essential for growth and positively influences NIMA function in Aspergillus nidulans. *J. Biol. Chem.* **2004**, *279*, 32373–32384.

(2) Lu, K. P.; Hanes, D.; Hunter, T. A human peptidyl-prolyl isomerase essential for regulation of mitosis. *Nature* **1996**, *380*, 544–547.

(3) Wulf, G. M.; Ryo, A.; Wulf, G. G.; Lee, S. W.; Niu, T.; Petkova, V.; Lu, K. P. Pin1 is overexpressed in breast cancer and cooperates with Ras signaling in increasing the transcriptional activity of c-Jun towards cyclin D1. *EMBO J.* **2001**, *20*, 3459–3472.

(4) Lu, K. P. Pinning down cell signalling, cancer and Alzheimer's disease. *Trends Biochem. Sci.* **2004**, *29*, 200–209.

(5) Bayer, E.; Goettsch, S.; Mueller, J. W.; Griewel, B.; Guiberman, E.; Mayr, L. M.; Bayer, P. Structural analysis of the mitotic regulator hPin1 in solution: Insights into domain architecture and substrate binding. *J. Biol. Chem.* **2003**, *278*, 26183–26193.

(6) Namanja, A. T.; Peng, T.; Zintsmaster, J. S.; Elson, A. C.; Shakour, M. G.; Peng, J. W. Substrate recognition reduces side-chain flexibility for conserved hydrophobic residues in human Pin1. *Structure* **2007**, *15*, 313–327.

(7) Ng, C. A.; Kato, Y.; Tanokura, M.; Brownlee, R. T. C. Structural characterisation of PinA WW Domain and comparison with other Group IV WW Domains, Pin1 and Ess1. *Biochim. Biophys. Acta* **2008**, *1784*, 1208–1214.

(8) Li, Z.; Li, H.; Devasahayam, G.; Gemmill, T.; Chaturvedi, V.; Hanes, S. D.; Van Roey, P. The structure of the Candida albicans Ess1 prolyl isomerase reveals a well-ordered linker that restricts domain mobility. *Biochemistry* **2005**, *44*, 6180–6189.

(9) Ranganathan, R.; Lu, K. P.; Hunter, T.; Noel, J. P. Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. *Cell* **1997**, *89*, 875–886.

(10) Verdecia, M. A.; Bowman, M. E.; Lu, K. P.; Hunter, T.; Noel, J. P. Structural basis for phosphoserine-proline recognition

2896 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Ng et al.

by group IV WW domains. *Nat. Struct. Biol.* **2000**, *7*, 639–643.

(11) Lu, K. P.; Liou, Y. C.; Zhou, X. Z. Pinning down proline-directed phosphorylation signaling. *Trends Cell Biol.* **2002**, *12*, 164–172.

(12) Macias, M. J.; Hyvönen, M.; Baraldi, E.; Schultz, J.; Sudol, M.; Saraste, M.; Oschkinat, H. Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature* **1996**, *382*, 646–649.

(13) Espanel, X.; Sudol, M. A single point mutation in a group I WW domain shifts its specificity to that of group II WW domains. *J. Biol. Chem.* **1999**, *274*, 17284–17289.

(14) Lu, P. J.; Zhou, X. Z.; Shen, M.; Lu, K. P. Function of WW domains as phosphoserine- or phosphothreonine-binding modules. *Science* **1999b**, *283*, 1325–1328.

(15) Sekerina, E.; Rahfeld, J. U.; Müller, J.; Fanghï; Anel, J.; Rascher, C.; Fischer, G.; Bayer, P. NMR solution structure of hPar14 reveals similarity to the peptidyl prolyl cis/trans isomerase domain of the mitotic regulator hPin1 but indicates a different functionality of the protein. *J. Mol. Biol.* **2000**, *301*, 1003–1017.

(16) Joseph, J. D.; Yeh, E. S.; Swenson, K. I.; Means, A. R.; Winkler, K. E. The peptidyl-prolyl isomerase Pin1. *Prog. Cell Cycle Res.* **2003**, *5*, 477–487.

(17) Wintjens, R.; Wieruxzeski, J.; Drobecq, H.; Rousselot-Pailley, P.; Buee, L.; Lippens, G.; Landrieu, I. 1H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides. *J. Biol. Chem.* **2001**, *276*, 25150–25156.

(18) Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendel, G.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W. Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10648–10653.

(19) Kato, Y.; Ito, M.; Kawai, K.; Nagata, K.; Tanokura, M. Determinants of ligand specificity in groups I and IV WW domains as studied by surface plasmon resonance and model building. *J. Biol. Chem.* **2002**, *277*, 10173–10177.

(20) May, A.; Sieker, F.; Zacharias, M. How to efficiently include receptor flexibility during computational docking. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 143–153.

(21) Lee, K.; Lee, J. W. Computational approaches to protein-protein docking. *Curr. Proteomics* **2008**, *5*, 10–19.

(22) Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J. Am. Chem. Soc.* **2006**, *128*, 2812–2813.

(23) *Maestro, version 8*; Schrodinger, LLC: New York, NY, 2007.

(24) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, W.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(25) Homeyer, N.; Horn, A. H. C.; Lanig, H.; Sticht, H. AMBER force field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine and phosphohistidine. *J. Mol. Model.* **2006**, *12*, 281–289.

(26) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(27) Case, D. A. et al. *AMBER 9*; University of California: San Francisco, CA, 2006.

(28) Chandrasekhar, S. Stochastic problems in Physics and Astronomy. *Rev. Mod. Phys* **1943**, *15*, 1–89.

(29) 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.

(30) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. A. Vulnerability in popular molecular dynamics packages concerning Langevin and Andersen dynamics. *J. Chem. Theory Comput.* **2008**, *4*, 1669–1680.

(31) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(32) Humphrey, W.; Dalke, A.; Schulten, K. VMD - visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38, 27–28.

(33) 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.

(34) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham III, T. E. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(35) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA helices. *J. Am. Chem. Soc.* **1998**, *120* (37), 9401–9409.

(36) Onufirev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem.* **2000**, *104*, 3712–3720.

(37) Lei, H.; Duan, Y. Two-stage Folding of HP-35 from Ab Initio Simulations. *J. Mol. Biol.* **2007**, *370*, 196–206.

(38) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-Protein Complex Ras-Raf. *J. Comput. Chem.* **2004**, *25*, 238–250.

(39) Kuhn, B.; Kollman, P. A. Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models. *J. Med. Chem.* **2000**, *43*, 3786–3791.

(40) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, *123*, 5221–5230.

(41) Hendsch, Z. S.; Tidor, B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **1994**, *3* (2), 211–226.

(42) Hendsch, Z. S.; Tidor, B. Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.* **1999**, *8* (7), 1381–1392.

(43) Sheinerman, F. B.; Honig, B. On the Role of Electrostatic Interactions in the Design of Protein-Protein Interfaces. *J. Mol. Biol.* **2002**, *318*, 161–177.

(44) Schapira, M.; Totrov, M.; Abagyan, R. Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit.* **1999**, *12* (3), 177–190.

(45) Borchert, T. V.; Kishan, K. V. R.; Zeelen, J. P.; Schliebs, W.; Thanki, N.; Abagyan, R.; Jaenicke, R.; Wierenga, R. K. Three new crystal structures of point mutation variants of mono TIM: conformational flexibility of loop-1, loop-4 and loop-8. *Structure* **1995**, *3* (7), 669–679.

(46) Gershenson, A.; Schauerte, J. A.; Giver, L.; Arnold, F. H. Tryptophan Phosphorescence Study of Enzyme Flexibility and Unfolding in Laboratory-Evolved Thermostable Esterases. *Biochemistry* **2000**, *39*, 4658–4665.

(47) Muralidhara, B. K.; Chen, M.; Ma, J.; Wittung-Stafshede, P. Effect of Inorganic Phosphate on FMN Binding and Loop Flexibility in Desulfovibrio desulfuricans Apo-flavodoxin. *J. Mol. Biol.* **2005**, *349*, 89–97.

(48) Lu, P. J.; Zhou, X. Z.; Liou, Y. C.; Noel, J. P.; Lu, K. P. Critical role of WW domain phosphorylation in regulating phosphoserine binding activity and Pin1 function. *J. Biol. Chem.* **2002**, *277*, 2381–2384.

(49) Zarrinpar, A.; Lim, W. A. Converging on proline: the mechanism of WW domain peptide recognition. *Nat. Struct. Biol.* **2000**, *7*, 611–613.

(50) Wu, B.; F., R. M.; Wei, J.; Yuan, H.; Dahl, R.; Zhang, Z.; Pellecchia, M. Discovery and Binding Studies on a Series of Novel Pin1 Ligands. *Chem. Biol. Drug. Des.* **2009**, *73*, 369–379.

JCTC Journal of Chemical Theory and Computation

Understanding Rubredoxin Redox Potentials: Role of H-Bonds on Model Complexes

Ana Patricia Gámiz-Hernández, Artur S. Galstyan, and Ernst-Walter Knapp*

Institute of Chemistry and Biochemistry, Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Fabeckstrasse 36a, D-14195 Berlin, Germany

Abstract: The energetics of redox states in different models of rubredoxin-like iron−sulfur complexes (ISC) was computed using a combination of density functional and electrostatic continuum theories. In agreement with experiment, the calculated redox potential for the small ISC model $[Fe(SCH_2CH_3)_4]^{1-,2-}$ in acetonitrile was −813 mV [Galstyan, A. S.; Knapp, E. W. *J. Comput. Chem.* **2009,** *30,* 203−211] as compared to the measured value of −838 mV. Surprisingly the experimental values for rubredoxin (Rd) are much higher ranging between −87 and +39 mV. These large variations in redox potentials of ISC models and ISC in Rd are due to specific conformational symmetries adopted by the ligands due to both the protein environment and type and the number of H-bonds, and the dielectric environment. In a dielectric environment corresponding to proteins ($\varepsilon = 20$), the computed ISC redox potentials shift positive by about 64 mV for Fe−S⋯H−N and 95 mV for Fe−S⋯H−O H-bonds, correlating well with data estimated from experiments on ISC proteins. In aqueous solutions ($\varepsilon = 80$), a positive shift of 58 mV was computed for Fe−S⋯H−O H-bonds (using a model with the same ISC conformation as in Rd) in agreement with a measured value for Rd with partially solvent exposed ISC. The latter demonstrates the dependence of the ISC redox potentials on the environment (solvent or protein). For a model whose chemical composition is analog to the relevant part of ISC in a specific Rd, the computed redox potential of the model agrees with the measured value in Rd. This study allows to understand redox potential shifts for small ISC models and ISC in proteins.

Introduction

Iron−sulfur complexes (ISC) are common redox-active cofactors in proteins.[2−6] These ISC proteins mediate electron transfer processes occurring in enzymatic catalysis, photosynthesis and respiration. ISC proteins are ubiquitous in nature and despite their similarity in physicochemical composition and structure of the iron sulfur complexes their redox potential values in proteins cover a large interval extending from −700 to +400 mV.[3,7−10] Interestingly, the measured redox potential of the most elementary ISC model $[Fe(SCH_2CH_3)_4]^{1-,2-}$ in acetonitrile (AN) is with −838 mV even lower.[11] This large variation in the ISC redox potentials was suggested to be due to differences in solvation,[12−15]

H-bond pattern[9,16−21] and the electrostatic environment of the redox centers provided by the protein.

There are two types of H-bonds commonly formed by ISC with ligating groups of the protein: Fe−S⋯H−N and Fe−S⋯H−O. The first type arises mainly from peptide backbone interactions with the ISC, while the second type is generally due to interactions with amino acid side chains of serine and tyrosine. Analysis of ISC models and ISC with peptide-based ligands involving Fe−S⋯H−N H-bonds[22−28] analogue to rubredoxin (Rd) and ISC in proteins[16,18,29−33] demonstrated that the redox potentials are shifted positively in the presence of H-bonds with sulfur. The influence of Fe−S⋯H−O H-bonds on the ISC redox potentials was quantified through mutations in corresponding proteins yielding a positive shift in Rd (65 mV for Ala to Ser mutation)[34] and Rieske proteins (44 or 65 mV for Tyr to Phe and −95 or −120 mV for Ser to Ala mutations).[35,36]

* Corresponding author: Fax: +493083856921. E-mail knapp@chemie.fu-berlin.de.

To study the influence of Fe−S···H−N H-bonds in proteins is more difficult, since these interactions involve backbone nitrogen atoms that cannot be removed without distorting the protein structure significantly. Hunsicker-Wang[18] analyzed the measured redox potential shifts of Rieske and Rieske-type proteins with different H-bond patterns leading to estimates of 70 mV for Fe−S···H−N and 140 mV for Fe−S···H−O. An analogue analysis is not possible for Rd, since the six amide H-bonds are present in all known Rd crystal structures. In this case, the influence of H-bonds on Rd redox potentials was studied by site-directed mutagenesis, which varied the redox potential, modifying the strength but not the number of H-bonds with the ISC.[29−31,33,37] Unfortunately, these series of mutations involve also changes in the dielectric environment and have side effects in protein conformation,[10,34,38] making it difficult to extract the influence of Fe−S···H−N H-bonds on the Rd redox potentials faithfully.

Theoretical results on ISC[39−41] and Rd[9,42−45] redox potentials obtained with quantum chemical and electrostatic energy computations demonstrated the difficulty to predict small variations (100 mV) in redox potentials and the limitations using continuum dielectric models to reproduce and understand properties of ISC in different environments. In the protein, specific H-bonds may help to delocalize and stabilize excess negative charges, which may particularly occur for ISC in a reduced state. Hence, to describe ISC redox potentials of model compounds and in proteins, H-bonds with the ISC sulfur atoms should be considered explicitly.[39,40]

Recently, electrostatic energy computations were combined with density functional theory (DFT) using a new functional B4(XQ3)LYP to evaluate accurate redox potentials of transition metal complexes in different solvents.[1] This method is able to predict experimental redox potentials of mononuclear transition metal (Fe, Ni, Mn) complexes in different solvents (water, AN, and dimethylformamide) with nearly 60 mV absolute value accuracy and is, therefore, an ideal tool to investigate how ISC redox potentials vary with different H-bond patterns. Similar accuracy was achieved earlier[46] for the redox potential calculations on the small organic compounds using the much more expensive ab initio G3MP2-method[47] combined with independent electrostatic computations.

In order to avoid the complications of ferromagnetic coupling observed in multinuclear ISC,[39,48−52] we have chosen the ISC of Rd that is one of the smallest redox-active proteins, containing the simplest ISC consisting of one iron coordinated by sulfur atoms of four cysteines. There are several crystal structures of Rd available (Figure 1a, for references see Supporting Information, Table S1.1).

The simplicity of ISC in Rd made it possible to construct several Rd-like ISC models where one can vary conformations and H-bond numbers and types and construct large models that include all structural elements to reproduce measured redox potentials of ISC models and Rd. In this study, model **4**, described later in detail, (Figure 1b) is most similar to the ISC in Rd. Therefore, it can best be used to
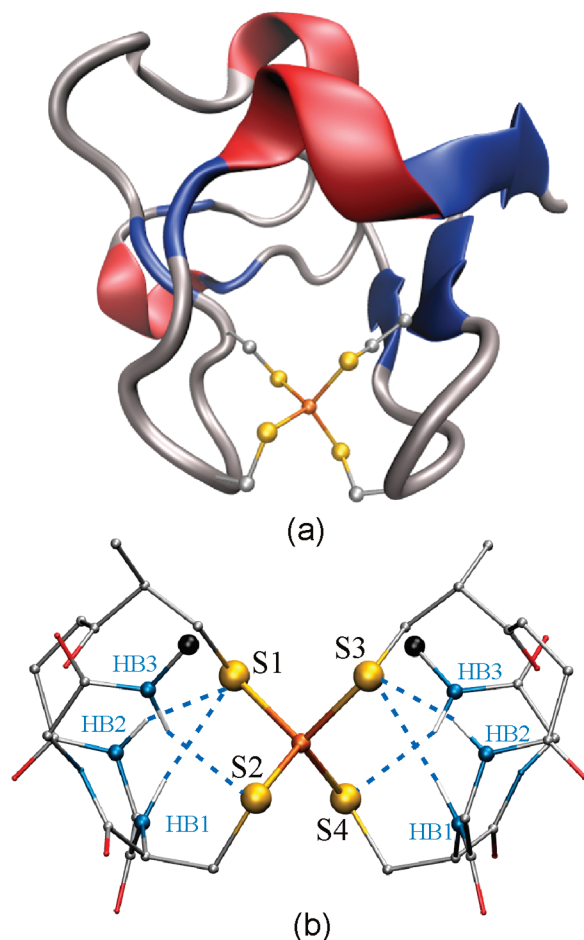

Figure 1. (a) The ISC in Rd is formed by one iron ligated by four Cys residues. The ISC in Rd interacts with the protein backbone through six amide H-bonds that are described by model **4** [depicted in part (b)], which mimics the three H-bond types (HB1, HB2, HB3) found in Rd without the sequence specific amino acid side chains. For a definition of the H-bond types, see Scheme 1 and text. It is the most detailed model to study the ISC in Rd. The redox potential calculated for this model in a dielectric continuum of $\varepsilon = 20$, corresponding to a protein environment, is +57 mV, which is close to values measured in different Rds (−87 to +39 mV). The color code of atoms corresponds to conventional CPK. We denoted in black color the two C-atoms, where model **4** was truncated from the ISC embedded in Rd.

study the dependence of Rd redox potentials on various details of the protein structure.

In this study we report on computed ISC redox potentials obtained with a combination of DFT and electrostatics[1,54] to estimate the effect of the two different H-bond partners (Fe−S···H−N and Fe−S···H−O) in ISC models and proteins. In agreement with experiments in proteins,[18,33−36] our computed ISC redox potentials are shifted positive by H-bonds to sulfur with a smaller influence for Fe−S···H−N than for Fe−S···H−O H-bonds due to the lower polarity of the NH versus the OH groups. We also found a linear correlation between the number of H-bonds and the ISC redox potential shift in agreement with experimental observations.[18] Another important effect influencing ISC redox potentials found in our computations is the closeness of the solvent (degree of solvent exposure) to the iron−sulfur core,

Scheme 1. The H-bonds of the ISC in *Clostridium pasterianum* (*Cp*) Rd (PDB code 1IRO[53]) Formed with the Protein Backbone NH-Groups Are Displayed[a]
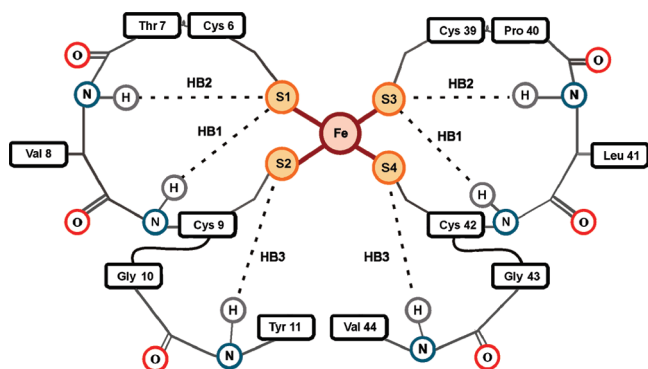
[a] There are six amide H-bonds of three different types related by two-fold rotation symmetry (C_2). Each H-bond type possesses different orientation and flexibility that affects the electronic state of the ISC model. The H-bonds of type HB1, HB2, and HB3 involve the residue pairs Cys9/Cys42, Val8/Leu41, and Tyr11/Val44, respectively. The two symmetry-related sulfurs from Cys6 and Cys39 are simultaneously involved in H-bonds of types HB1 and HB2, while the two sulfurs from Cys9 and Cys42 form only a single H-bond of type HB3. Further explanation of the H-bond types is given in the text. In Rd, the orientations of Cys6 and Cys9 lead to C_2-symmetry in contrast to the S_4-symmetry observed in small ISC models. The backbones of Cys6 and Cys39 are buried in the protein, while the backbones of Val8/Cys9/Tyr11 and Leu41/Cys42/Val44 are solvent exposed.

which varies with the size of the ISC model. These findings are relevant for ISC proteins where we observed a strong dependence of the ISC redox potential with the dielectric environment,[14,15] which combined with H-bond effects provide the key to understand the variation of the redox potentials for different ISC models and ISC in proteins.

Materials and Methods

ISC H-Bonds in Rd. The ISC redox potentials in Rd are affected by the strength of the six amide H-bonds formed with the protein backbone.[16,29–33] We distinguish three types of H-bonds (Figure 1a) (HB1, HB2, HB3, depicted in Scheme 1). Each of the ISC sulfur atoms, S1 and S3, is engaged in two H-bonds, which we call type HB1 and HB2. The two HB1 are formed between the NH-groups of the other two Cys ligating to the iron with sulfur atoms S2 and S4. The two HB2 are formed with the NH-groups of the residues next to these Cys in the N-terminal direction. The ISC sulfur atoms S2 and S4 are involved in single H-bonds each, called type HB3. These H-bonds are formed with the amide backbone of the residues two positions in the C-terminal direction from the ligating Cys providing the sulfur involved in this H-bond. The amide groups forming HB1 are more constrained than that of HB2 and HB3, since they belong to the backbone of the ligating Cys.

H-bonds of Rd-like ISCs. Model **1a** with redox states $[Fe(SEt)_4]^{1-/2-}$ is the most simple Rd-like ISC (Figure 2). Its initial coordinates were taken from a synthetic ISC models[55] of known crystal structure [Cambridge structural database (CSD) reference code CANDAW] possessing a nearly ideal S_4-symmetry. Model **2a** has the same chemical
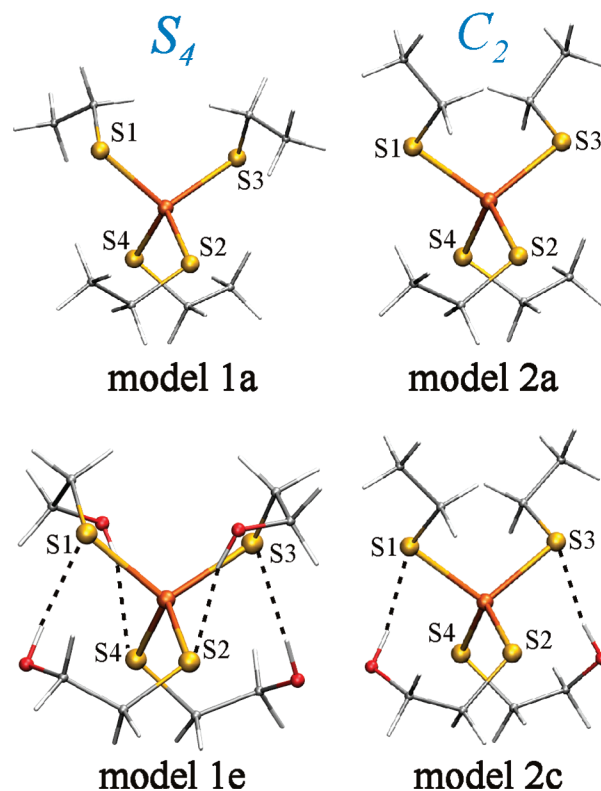

Figure 2. ISC models with symmetry S_4 and C_2 (models **1a** and **2a**, respectively) and ISC models **1e** and **2c** with Fe−S···H−O H-bonds derived from models **1a** and **2a**, respectively. The H-bond models **1b**−**e** and **2b**−**c** were constructed from model **1a** and **2a**, respectively, by substituting the corresponding terminal −CH$_3$ groups by −CH$_2$OH groups, which form intramolecular H-bonds with the sulfur atoms and preserve the same solvation surface to account for the redox potential shift dependence on the number of H-bonds. Due to symmetry, model **1a** allows formation of a maximum four H-bonds (model **1e**), while in model **2a** just two H-bonds are possible (model **2c**).

composition as model **1a** but was built in order to represent a conformation closer to the one found in Rd crystal structures, where the ISC possesses a distorted C_2-symmetry due to steric constrains.[56] The initial geometry of model **2a** is based on the coordinates of the Fe(Cys)$_4$ complex from *Clostridium pasterianum* (*Cp*) Rd (PDB code 1IRO[53]), where the backbone of each Cys was substituted by a methyl group resulting in sulfur with an ethyl group (Figure 2). The use of ISC with ethyl groups was recommended by Szilagy and Winslow,[57] to avoid the spread of negative charge in the H-atoms observed in ISC models with methyl groups.[33,39,50,58] Models **1a** and **2a** were subsequently used to model systematically ISCs with a different H-bond pattern, replacing the terminal −CH$_3$ groups of the ligands by one to four −CH$_2$OH groups in model **1a** or one to two −CH$_2$OH groups in model **2a** to form intramolecular S···H−O H-bonds yielding models **1b**−**e** and **2b**−**c** with S_4- and C_2-based symmetry, respectively (listed in Table 3). ISC model complexes with S···H−O H-bonds were built to mimic the interaction of a Ser side chain with ISC observed in several ISC proteins.[18,34,59] Due to the C_2-symmetry of model **2a**, it was only possible to consider ISCs with one and two S···H−O H-bonds.
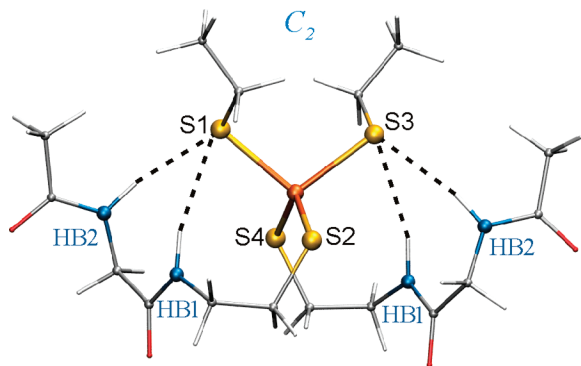
Figure 3. ISC model **3a** with four amide H-bonds (two HB1 and two HB2, as defined in Scheme 1). ISC models related to model **3a** but with a smaller number of H-bonds [models **3b–h**] were obtained by substituting one (**3b**, **3c**), two (**3d**, **3e**), three (**3f**, **3g**), or all four NH-groups by CH$_2$-groups. Thus, the solvent boundaries of the ISC models are approximately preserved. This avoids influences from different electrostatic boundaries on the ISC redox potentials and allows to observe how the redox potential shifts with the number of H-bonds.

Model **3a** (Figure 3) possesses four H-bonds with NH-groups two of type HB1 and two of type HB2 reflecting part of the H-bond pattern of the ISC in Rd (see Scheme 1). The coordinates of this model were generated based on the crystal structure of *Cp* Rd.[53] Models **3b–h** were derived from model **3a** by substituting one (**3b**, **3c**), two (**3d**, **3e**), three (**3f**, **3g**), or four (**3h**) of the NH-groups by CH$_2$-groups to reduce the number of H-bonds accordingly.

Model **4** (Figure 1b) includes all six amide H-bonds formed with the protein backbone as in Rd (Scheme 1). The coordinates of this model were modeled based on the crystal structure of *Cp* Rd.[53] For the backbone parts of the ISC model **4**, which are not engaged in H-bonds with sulfur and not relevant for this study, we substituted $-N-H$ and $-C=O$ groups by $-CH_2$ groups without distorting the remaining backbone conformation. This, relatively large ISC model, is a trade-off between the limitations imposed by the performance of quantum chemical computations and the sufficiently faithful modeling of the ISC protein environment. Nevertheless, it can account for three main factors that shift the redox potential in ISC: ISC symmetry, H-bonds, and dielectric environment.

Redox Potentials. Absolute values of the ISC redox potentials were calculated using the B4(XQ3)LYP approach,[1] which employs the DFT method with a modified functional combined with a postcomputational charge dependent empirical DFT correction and subsequent optimized electrostatics to compute solvation energies. In the B4(XQ3)LYP approach, we compute the free energy difference of reduced minus oxidized state in vacuum ΔG_g°, combining quantum chemical DFT-based ($G_g^{\circ, \text{B4LYP}}$) and empirical ($G_X$) contributions as follows:

$$\Delta G_g^\circ = \Delta E_0^{\text{B4LYP}} + \Delta \text{ZPE} + \Delta G_{0 \to 298K} + \Delta G_X(q_{\text{red}}) =$$
$$\Delta G_g^{\circ, \text{B4LYP}} + \Delta G_X(q_{\text{red}}) \quad (1)$$

with the empirical term ΔG_X (in units of kcal·mol^{-1}), depending on the total charge q_{red} of the reduced state, where $\Delta G_X(q_{\text{red}}) = G_X(q_{\text{red}}) - G_X(q_{\text{red}} + 1)$ and

$$G_X(q) = -0.333q^3 + 1.545q^2 + 21.634q \quad (2)$$

The B4LYP functional contains the same terms as the B3LYP but with weights that differ for the exact exchange. The weight parameters for the two exchange terms in the B4LYP functional are decoupled. While the parameter for the local exchange remains at the B3LYP value of 0.80, the exact exchange term is reduced from 0.20 to 0.12, and the other parameters are the same as in the B3LYP functional.[1] An empirical correction G_X compensates for the energy deficit caused by the decrease of the exact exchange contribution and is fitted after the SCF computation to reproduce correctly the experimental redox potentials.[1]

The quantum chemical ground-state electronic energies (E_0^{B4LYP}) are computed in vacuum for both the oxidized and reduced states using the incomplete DFT functional B4LYP,[1] while the zero-point vibrational energies for the vibrational ground state (ZPE) and the free energies for the thermal excited states at 298 K ($G_{0 \to 298\ K}$) are calculated with the standard B3LYP functional according to the previous work.[1]

The free energy of a redox reaction in solution ΔG_s° can be computed as

$$\Delta G_s^\circ = \Delta G_g^\circ + \Delta \Delta G_{\text{sol}}^\circ \quad (3)$$

where ΔG_g° and $\Delta \Delta G_{\text{sol}}^\circ$ are, respectively, the differences of vacuum and solvation energies between the reduced and oxidized states.

The standard (°) redox potential E° of the redox reaction $\text{Ox}_s + e_g \to \text{Red}_s$ in solution relative to the absolute potential of the standard hydrogen electrode (SHE), E_{SHE}°, is given by

$$E^\circ = E_{\text{SHE}}^\circ - \frac{\Delta G_s^\circ}{nF} \quad (4)$$

where $F = 23.06$ kcal·mol^{-1} V^{-1} is the Faraday constant and n is the number of transferred electrons (in our case $n = 1$).

For the E_{SHE}°, the values of 4.43,[61] 4.44,[62] and 4.36[63] are known. Since the value of the free energy of hydration, included in the calculation of E_{SHE}° stems from the indirect measurements, their accuracy will influence the value of E_{SHE}°. The value of -4.36 eV for E_{SHE}° has been suggested recently,[63] based on the newest results of proton hydration energy from gas-phase experiments.[64] However, the exact value of E_{SHE}° is not crucial in our method, since the computed redox potentials can be easily adjusted to another value of E_{SHE}°. Here, we use -4.44 V,[62] as recommended by IUPAC.

Quantum Chemical Computations. Quantum chemical calculations were performed using DFT. Geometries of all considered ISC models were optimized in vacuum using the B3LYP[65–69] functional with LACVP[70] effective core potential for the iron atom and the 6-31G** basis functions for main-group atoms (LACVP**). Single-point electronic energies E_0^{B4LYP} needed for the computation of absolute redox potentials of the model complexes within the B4(XQ3)LYP approach[1] were calculated for optimized geometries using the incomplete functional B4LYP[1] with the basis sets LACV3P**++, including the LACV3P[71] effective core potential for the iron atom and the 6-311++G** basis

function for all other atoms. Vibrational frequency calculations were done using the B3LYP functional with LACVP** basis set without rescaling of vibrational frequencies. Since the ISCs in Rd are in the high spin state,[72] the model compounds with total spins $S = 5/2$ and $S = 2$ for reduced and oxidized forms, respectively, were considered. All quantum chemical computations were performed by the program Jaguar 5.5.[71]

Atomic Partial Charges. Atomic partial charges of the ISC were obtained from electrostatic potentials using the restraint electrostatic potential (RESP) method.[73,74] The electrostatic potentials were calculated quantum chemically in vacuum at the B3LYP/LACVP** level using Jaguar 5.5.[71] The atomic partial charges were determined in two stages, using hyperbolic restraints with the total charge fixed. In the first stage of the RESP procedure, the atomic charges were allowed to change with a restraining weight of 0.0005 au. In the second stage, the charges on C- and H-atoms were left free, while all other atomic charges were constrained at their values obtained from the first stage, using a restraining weight of 0.001 au.

Solvation Energy. For the computation of solvation energies ΔG_{sol}°, the environment was represented as a dielectric continuum. For water, a high dielectric constant ($\varepsilon_W = 80.0$) was used, while for AN a lower dielectric constant ($\varepsilon_{AN} = 37.5$) was used. The value of the dielectric constant appropriate to describe proteins depends on the atomic detail of the protein model[75,76] and the method used to represent the continuum[77] and can vary typically between 4 and 30. For protein models with all atomic details, $\varepsilon_P = 4$ is generally used.[78] For implicit protein models with no atomic detail, the appropriate dielectric constant varies typically between 10 and 30. For the sake of simplicity, we consider in this application an implicit protein model with no explicit atom representation assuming $\varepsilon_P = 20$.

The solvation energies were computed with the program module "solvate" from the MEAD program suite.[79,80] The dielectric constant within the solute cavity was taken to be $\varepsilon = 1$. The solute cavity was defined by the joint van der Waals volumes of all atoms using atomic radii, as in a previous work.[1] Since H-atoms have the ability to approach other atoms most closely, they will have the strongest electrostatic interactions with other atoms, if they are polar. To account for this effect, we use larger atomic radii for solute atoms in aprotic solvents, like AN, than in protic solvents, since in protic solvents polar hydrogens are absent.[46] Outside the solute cavity, the dielectric constant appropriate for the solvent/protein was taken, and the ionic strength was set to zero. As for the solvent probe radius, we used 1.4 Å for water and molecular groups from the protein environment, while we used 2.23 Å for AN.[1] A two-step focusing procedure was used to solve the Poisson equation numerically on a grid consisting of $(189)^3$ points, using first a low- and then a high-resolution grid with lattice constants of 0.4 and 0.1 Å, respectively.

Results and Discussion

The shifts of redox potentials of our ISC models depend basically on the following factors: the change of symmetry

of the ISC (from S_4 to C_2) and the number and type of H-bond partners (Fe−S···H−O or Fe−S···H−N) interacting with sulfur ligated to iron and dielectric environment (degree of solvent exposure).[12−19] These factors are presented and discussed in detail in the following sections. We start with the symmetry effect, discussing also how it affects the charge distribution. Then we analyze the influence of the H-bond partners for different numbers and types of H-bonds. Finally, with a model similar to ISC in rubredoxin (Figure 1b), we quantify the factors shifting the redox potential from the ISC models to protein.

ISC Models with S_4 and C_2 Symmetry. Geometry comparison of an ISC model with similar composition but different conformational symmetry of the ligands (model **1a** with S_4 and model **2a** with C_2 symmetry, Figure 2) shows that the ISC core structure consisting of Fe(SC)$_4$ is nearly invariant (see Supporting Information, Table S3.1). Structural differences appear in the dihedral angles Fe−S−C−C involving the ethyl groups, which in the oxidized state are at 8° for Fe−S1−C−C and Fe−S3−C−C and at −8° for Fe−S2−C−C and Fe−S4−C−C with S_4 symmetry, while in the model with C_2 symmetry the corresponding values are 17° and 85°, respectively (see Figure 2).

The atomic partial charges of the iron and sulfur atoms of models **1a** and **2a** essentially do not depend on symmetry, while the charges of the attached ethyl groups do depend on symmetry (Supporting Information, Table S4.1). The groups (S2/S4)−CH$_2$−CH$_3$ of model **1a** carry larger dipole moments than the corresponding atoms of model **2a** with C_2 symmetry.

Table 1 shows the energies of the most elementary ISC models **1a** and **2a** with S_4 and C_2 symmetry, respectively. The energies are given in kcal/mol relative to the vacuum oxidized state in S_4 symmetry (whose energy was set to zero) using the conventional B3LYP DFT functional as well as the alternative approach with the B4(XQ3)LYP DFT functional and subsequent post-SCF correction.[1] Comparing the vacuum energies, we recognize that the ISC model in the oxidized state is more stable in the S_4 than in the C_2 symmetry for both DFT functionals, while this is opposite for the reduced state. This behavior changes in the condensed phase where, according to our computations, the S_4 symmetry conformer is more stable in both redox states (valid for environments with $\varepsilon > 10$). We have tested this behavior for both water and acetonitrile. In Table 1 we show the energies for the dielectric constant of $\varepsilon = 37.5$, corresponding to acetonitrile but used the small set of atomic radii suitable of protic solvents, while AN is an aprotic solvent. This setting is close to the electrostatics prevailing in the protein environment with an effective dielectric constant of $\varepsilon = 20$, where we also verified that the S_4 conformer (model **1a**) is more stable than the C_2 conformer (model **2a**) in both redox states. Hence, the fact that in the protein the ISC possesses C_2 symmetry is due to the constraints, which the protein applies to the ISC ligand geometries.

Table 2 shows the redox potentials of models **1a** and **2a.** The calculated redox potential of model **1a** in AN is −813 mV,[1] which is in agreement with the measured value of −838 mV.[11] This very negative redox potential demonstrates the

Table 1. Computed Energies (kcal/mol) of the ISC Models **1a** and **2a** (Figure 2) Relative to the Oxidized States in Vacuum Using the DFT Approach with Functionals B3LYP and B4(XQ3)LYP[1] and Symmetries S_4 and C_2

	symmetries			
ISC models	$S_4{}^a$	$C_2{}^b$	$S_4{}^a$	$C_2{}^b$
Energies in Vacuum, ΔG_g^o (eq 1)				
	B3LYP		B4(XQ3)LYP[1]	
$[Fe(SCH_2CH_3)_4]^{1-}$	0.00^c	3.40	0.00^c	3.63
$[Fe(SCH_2CH_3)_4]^{2-}$	31.54	28.92	29.09	26.61
Energies in Acetonitrile^d ΔG_s^o (eq 3)				
	B3LYP		B4(XQ3)LYP	
$[Fe(SCH_2CH_3)_4]^{1-}$	−44.33	−39.55	−44.33	−39.32
$[Fe(SCH_2CH_3)_4]^{2-}$	−126.31	−125.52	−128.75	−127.82

^{a} Model **1a**. ^{b} Model **2a** (see Figure 2). ^{c} To facilitate comparison of energies for different redox states and symmetries, the energy of the oxidized state in symmetry S_4 was set to zero. ^{d} $\varepsilon = 37.5$ corresponds to AN, but with smaller solute atomic radii it corresponds to protic solvent environment like water or protein.

Table 2. Computed Free Energy Differences and Comparison of Calculated and Measured Redox Potentials of the ISC Model $[Fe(SCH_2CH_3)_4]^{1-/2-}$ in Two Different Symmetry States, S_4 (model **1a**) and C_2 (model **2b**)

	E_{comp}^o (eq 4), mV			E_{exp}^o, mV
ISC models	water^a	AN^a protic	AN^b aprotic	AN
$[Fe(SCH_2CH_3)_4]^{1-/2-}$ (S_4)	−15	−104	-813^c	-838^d
$[Fe(SCH_2CH_3)_4]^{1-/2-}$ (C_2)	56	−28	−602	–

^{a} Inside the ISC volume, the dielectric constant is unity, while outside we used 80 for water and 37.5 for acetonitrile (AN). Here, we use the same smaller solute atomic radii for AN as for water, corresponding to a protic rather than an aprotic solvent. ^{b} In this case we use the larger solute atomic radii corresponding to AN. ^{c} Ref 1. ^{d} Ref 11.

discrepancies in redox potentials between the small model compounds in aprotic solvents and the redox potentials reported in Rd, which is between −87 and 39 mV (see Supporting Information, Table S1). Changing the solvent from AN to water stabilizes the reduced state of model **1a** (S_4 symmetry), shifting the redox potential by a large positive value of 798 mV as observed in experiments of ISC measured in aprotic and aqueous solutions.[22,23] To demonstrate this effect more clearly, we computed the redox potential of ISC in a dielectric corresponding to AN but assumed that AN is a protic solvent where the small solute atomic radii apply. Assuming such an artificial solvent condition, the redox potentials of the ISC are computed to be −104 and −28 mV for S_4 (model 1**a**) and C_2 (model 2**a**) symmetry, respectively. This is an enormously large shift to more positive redox potentials as compared to the conditions appropriate for the aprotic solvent AN. These large shifts are due to the smaller solute atomic radii, which are applicable for protic dielectric environments of water and protein as compared to AN. In an environment of a large dielectric constant, smaller solute atomic radii lead implicitly to strong H-bonds, formed with the electronegative atoms of the solute.

In our study of Rd-like ISCs, we observed that the change of conformational symmetry of the ISC (from S_4 to C_2) is also up-shifting the redox potential. Due to structural peculiarities, the free energy differences (ΔG_g^o and $\Delta\Delta G_{sol}^o$, see eq 2) between the reduced and oxidized states are larger in the conformation with S_4 than with C_2 symmetry. This leads to an increase of the $[Fe(SCH_2CH_3)_4]^{1-/2-}$ redox potential by 71 mV in water and 211 mV in AN when changing the symmetry from S_4 to C_2 (Table 2). To analyze the symmetry related redox potential shifts more in detail, we computed the ISC redox potentials of model **1a** in water using the coordinates of S_4 symmetry combined with atomic partial charges of C_2 symmetry and vice versa yielding redox potentials of −142 and −273 mV, respectively. Comparing these values with the computed redox potentials combining coordinates and charges appropriately (Table 2), we observe a down-shift of the redox potentials for the artificial combinations of coordinates and charges, which is much

stronger for C_2 than for S_4 symmetry. Interestingly, changes in the charges and the coordinates shift the redox potentials in opposite directions such that the net effect in water is a moderate upshift of the ISC redox potential going from S_4 to C_2 symmetry. However, in AN, this upshift is considerably larger at about +200 mV. Hence, the variation in ISC redox potentials with symmetry clearly shows that redox potentials in Rd can be better understood, if the ISC models considered possess the same symmetry as in the protein.[56]

ISC Models with Fe−S···H−O H-Bonds. The considered ISC models with different H-bond geometries of type Fe−S···H−O are presented in Table 3. The ISC models **1(b−e)** and **2(b−c)** are variations of models **1a** and **2a**, respectively, which involve OH groups that form H-bonds with sulfur. The OH group H-bond partner is not present in native Rd but was created for a mutant of *Pyrococcus abyssi* (*Pa*)[34] to measure the influence of S···H−O H-bond in Rd. The crystal structure of the *Pa* mutant was solved at high resolution (0.86 Å), and the redox potential was estimated to shift by 65 mV due to the Fe−S···H−O H-bond.[34] In this crystal structure, the S···O distance is 3.24 Å, and the S···H−O angle with the H-bond partner is 158.7°. The geometry optimized structures of our ISC models yield in the oxidized state S···O distances varying from 3.26 to 3.58 Å and H-bond angles in the range 164.3 to 178.6°. Hence, our calculated S···O distances in the ISC models (Supporting Information, Table S3.3) seem to agree with the value obtained from the Rd crystal structure, although, the calculated H-bond angles are closer to linearity than in the crystal structure. The latter is actually the reason that the S···O distance in the crystal is smaller than the average distance obtained with the ISC models.

In proteins, the S···O distances of H-bonds between OH groups and sulfur of cysteines and methionines are 3.45 and 3.32 Å, respectively, with standard deviations of 0.2 Å.[81] Considering small models with these H-bonds, we obtained after geometry optimization, using the same level of theory as for the ISC, 3.39 and 3.45 Å for the S···O distance in Cys−hydroxyl and Met−hydroxyl, respectively (see Supporting Information, Table S3.6), which are quite similar to the corresponding distances in proteins. The small discrepancies between our computed values and the reported values in protein crystal structures may be related to the different types of constraints in our models and in proteins.[34]

2904 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Gámiz-Hernández et al.

Table 3. Computed Redox Potentials E^o_{comp} of ISC Models Based on S_4 (models **1(a–e)**) and C_2 (models **2(a–c)**) Symmetry, Depending on the Number of Fe–S···H–O H-Bonds

ISC models[a]	N_H[b]	E^o_{comp}, mV	
		$\varepsilon = 80$[c]	$\varepsilon = 20$[c]
1a $[Fe(SEt)_4]^{1-/2-}$ (S_4)	0	−15	−240
1b $[Fe(SEt)_3(SEtOH)]^{1-/2-}$ (C_1)	1	80	−138
1c $[Fe(SEt)_2(SEtOH)_2]^{1-/2-}$ (C_2)	2	143	−69
1d $[Fe(SEt)(SEtOH)_3]^{1-/2-}$ (C_1)	3	254	44
1e $[Fe(SEtOH)_4]^{1-/2-}$ (C_2)	4	350	144
2a $[Fe(SEt)_4]^{1-/2-}$ (C_2)	0	56	−165
2b $[Fe(SEt)_3(SEtOH)]^{1-/2-}$	1	70	−143
2c $[Fe(SEt)_2(SEtOH)_2]^{1-/2-}$ (C_2)	2	173	−36

[a] The symmetry types of the ISC models are given in parentheses. Ligand abbreviations: SEt = ethanethiol-1-ide and SEtOH = sulfanylmethanol-1-ide. [b] Number of (S···H–O) H-bonds. [c] Inside the ISC complex, the dielectric constant is unity, while for the environment, a dielectric constant of 80 corresponds to water and 20 corresponds to protein.

The redox potentials of the ISC models **1a** and **2a** are shifted positively, increasing the number of S···HO H-bonds (models **1(b–e)** and **2(b–c)**) (Table 3) depending linearly on the number of H-bonds, Figure 4. For the series of H-bonds in ISC models **1(a–e)**, the linear fit yields an increase per H-bond of 90 mV in water ($\varepsilon = 80$) and 95 mV in low dielectric medium, corresponding to a protein environment ($\varepsilon = 20$). Due to the C_2 symmetry (see Figure 2), the ISC model **2a** can form only up to two H-bonds of ligands with sulfur. For this compound, the dependence of the ISC redox potentials with the number of H-bonds is evidently not linear (Figure 4). The deviation from linearity is probably due to the single H-bond in ISC model **2b** that disturbs the C_2 symmetry, which is perfectly maintained for the corresponding ISC models with none or two H-bonds. The redox potential shift due to two H-bonds is for the ISC model **2c** $\Delta E = 117$ mV (58 mV for one H-bond), which is comparable to the redox potential shift of the corresponding ISC model **1c** (with two H-bonds in S_4 symmetry yielding the shift $\Delta E = +158$ mV, see Table 3). These results indicate that the influence of H-bonds on the ISC redox potential depends on the type of symmetry of the ISC, and in the series of models **1(a–e)**, the H-bonds have a stronger effect on the ISC redox potentials than in models **2(a–c)**.

The experimental shifts of ISC redox potentials estimated for the formation of Fe–S···H–O H-bonds in ISC proteins range from 44 to 120 mV.[34–36] Hence, our calculated shift of 90 mV for this type of H-bond is in the range of measured values. Deviations from measured redox potentials may be due to constraints from and interactions with the protein environment.

ISC Models with Fe–S···H–N H-Bonds. Table 4 and Figure 3 contain the values of computed ISC redox potentials of models **3(a–h)** with a variable number of amide H-bonds. In Scheme 1, corresponding to model **3a**, we depicted four intramolecular H-bonds of type HB1 and HB2, involving the Cys backbone amide group and the protein backbone as in Rd, respectively. Models **3(b–h)** are variations of model **3a**, possessing similar chemical composition and dielectric boundary but involving different numbers of intramolecular H-bonds.
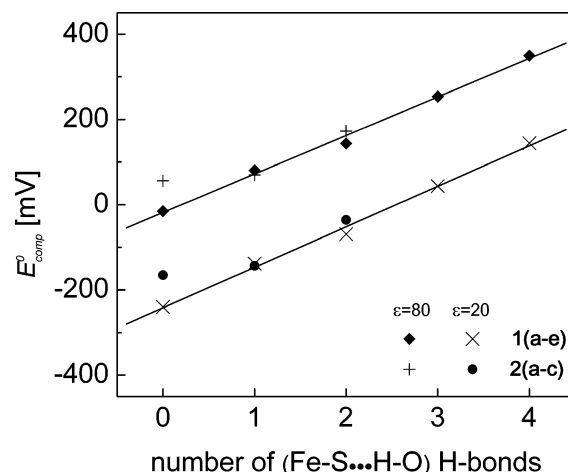
Figure 4. Calculated redox potentials as a function of the number of Fe–S···H–O H-bonds in ISC models in two dielectric environments: water- ($\varepsilon = 80$) and protein-like ($\varepsilon = 20$). Models based on S_4 symmetry allow a maximum of four H-bonds (model **1b–e**), while models based on C_2 symmetry allow a maximum of only two H-bonds (model **2a–c**).

Table 4. Computed Redox Potentials E^o_{comp} of ISC models **3(a–h)** with a different Number of Fe–S···H–N H-Bonds of types HB1 and HB2

ISC model[a]	N_{HB1}/N_{HB2}[b]	E^o_{comp}, mV	
		$\varepsilon = 80$[c]	$\varepsilon = 20$[c]
3a $[Fe(SEt)_2(saa)_2]^{1-/2-}$	2/2	6	−161
3b $[Fe(SEt)_2(saa)(sac)]^{1-/2-}$	2/1	−22	−199
3c $[Fe(SEt)_2(sac)(sca)]^{1-/2-}$	1/2	−91	−261
3d $[Fe(SEt)_2(sac)_2]^{1-/2-}$	2/0	−100	−278
3e $[Fe(SEt)_2(sca)_2]^{1-/2-}$	0/2	−58	−232
3f $[Fe(SEt)_2(sac)(scc)]^{1-/2-}$	1/0	−171	−350
3g $[Fe(SEt)_2(sca)(scc)]^{1-/2-}$	0/1	−180	−357
3h $[Fe(SEt)_2(scc)_2]^{1-/2-}$	0/0	−238	−419

[a] For ligand abbreviations see Supporting Information, Table S2.1. [b] N_{HB1} and N_{HB2} are the numbers of H-bonds of type HB1 and HB2, respectively, Scheme 1. [c] Inside the ISC complex, the dielectric constant is unity, while for the environment, the dielectric constant of 80 corresponds to water and 20 corresponds to protein.

The S···N distances obtained in our models after geometry optimization of crystal structures vary from 3.5 to 4.3 Å (Supporting Information, Table S3.4), being in general larger than 3.3–4.0 Å observed in Rd crystal structures. Analyzing H-bonds in proteins between the NH group of the backbone and the sulfur of Cys and Met, the S···N distances are 3.61 and 3.59 Å, respectively, with standard deviations of 0.2 Å.[81] The S···N distances obtained after quantum chemical geometry optimization of Cys–amide and Met–amide models (using the same level the theory as for the ISC models) are 3.69 and 3.64 Å for Cys and Met, respectively (see Supporting Information, Table S3.6), which essentially agree with corresponding distances in proteins. Hence, constraints by the protein environment are likely not the reason for discrepancies between the computed and the measured S···N distances in ISC models and Rd. The reasons that we computed larger S···N distances in our ISC models as compared to the values in the Rd crystal structures could be the specific constraints of the ISC models not present in
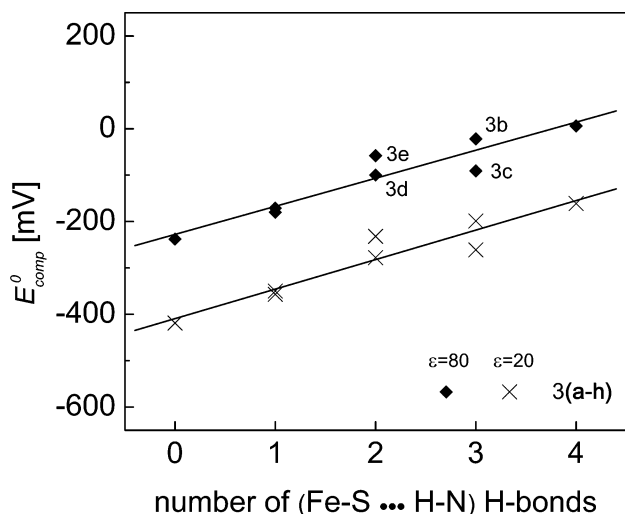
Redox Potentials of Rubredoxin Model Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2905**

Figure 5. Calculated redox potentials as a function of the number of (S···H−N) H-bonds in the ISC models of type HB1 and HB2 (see Scheme 1) in water ($\varepsilon = 80$) and protein environments ($\varepsilon = 20$). The ISC model pairs (**3b,c**), (**3d,e**), (**3f,g**) possess the same number of H-bonds in different positions (see Figure 3).

the protein or the level of quantum chemical theory,[57] which may not be appropriate in the presence of iron.

The redox potentials of ISC models **3(a−h)** with S···H−N H-bonds are plotted in Figure 5. The linear fit yields a shift of the redox potential per H-bond of 60 and 65 mV in water and protein environments, respectively. This value agrees with the estimated value of 70 mV from measurements in Rieske and Rieske-type proteins.[18] From our ISC models, it is not possible to distinguish which type of H-bond (HB1 or HB2) exerts a stronger redox potential shift.

Rubredoxin Redox Potential. Model **4** is our largest ISC model (101 atoms) in this study designed to mimic the H-bond pattern in Rd. Although this model follows the orientation of the ISC and H-bonds as found in *Cp* Rd, it does not include the side chains denoted in Scheme 1. The ISC model **4** corresponds to a sequence involving only glycine (G) at two positions with an incomplete backbone (IBB) [8(G)10(G)11(IBB)41(G)43(G)44(IBB)]. It can best be compared with the *Cp* Rd mutant[29] [8(G)10(G)11(Y)41(L)-43(G)44(G)]. Although there is an experimental value of the redox potential reported for this *Cp* Rd mutant (+39 mV),[29] there is no crystal structure for this protein available. We compared model **4** in the oxidized state with the wild-type *Cp* Rd crystal structure (PDB code 1IRO), yielding a rmsd (without hydrogen atoms) of 0.577 Å. Major structural deviations of model **4** compared to the ISC in Rd are with respect to the two terminal methyl groups (black carbon atoms in Figure 1b), corresponding to the IBBs at positions 11 and 44, which are constraints in the protein. We also observed deviations in the six dihedral angles H−N−C−O, whose hydrogens are involved in the H-bonds.

Comparing the reduced and oxidized state geometries, the ISC model **4** yields an rmsd of 0.25 Å, that is comparable to 0.18 Å obtained by considering the equivalent atoms in the crystal structures of *Cp* Rd.[82] We can also compare equivalent atoms in the optimized geometries of model **4**

Table 5. Calculated Redox Potentials of ISC Model **4** in Different Dielectric Environments and Comparison with Experimental Values with Similar ISC Ligand Composition[a]

ISC model	E^0_{comp}, mV		
	$\varepsilon = 80^b$	$\varepsilon = 20^b$	AN[c]
4 [Fe(saaa)$_2$]$^{1-/2-}$ ^{d}	194	57	−100
		Rd interval[e]	Rd analogue[f]
measured redox potentials		(−87, +39)	−136

[a] Model **4** possesses six amide H-bonds similar to Rd, see Scheme 1 and Figure 1b. [b] Inside the ISC complex, the dielectric constant is unity, while for the environment, the dielectric constant of 80 corresponds to water and 20 corresponds to protein. [c] AN with $\varepsilon = 37.5$. [d] For the ligand abbreviation, see Supporting Information, Table S1 [e] Rd experimental values,[29] for a complete list see Supporting Information, Table S2. [f] Synthetic peptide-based ISC model as Rd analogue.[24]

with the *Cp* Rd crystal structures in the corresponding redox states: 1FHH (oxd) and 1FHM (red),[82] yielding an rmsd of 0.57 and 0.52 Å for the oxidized and reduced states, respectively. Major structural deviations of model **4** from the Rd crystal structures are with respect to the two terminal methyl groups (black carbon atoms in Figure 1b), corresponding to the IBBs at positions 11 and 44, which are constraints in the protein. We also observed deviations in equivalent dihedral angles H−N−C−O involving the H-bond hydrogens between the ISC model and the Rd. In the latter case, the H-bond strengths vary with the amino acid type in those positions.[33]

The redox potentials computed of ISC model **4** are shown in Table 5. For water ($\varepsilon = 80$), it is +194 mV, while for the protein environment ($\varepsilon = 20$), it is +57 mV. The measured redox potentials of Rd vary between −87 and +39 mV in different species of Rd[9,53,72,83,84] (see Supporting Information, Table S1.1 for a more complete list of references). Our computed ISC redox potential of model **4** is more positive, which can qualitatively be understood, since it involves only Gly residues in positions with amide H-bonds with sulfur. Experiments with *Cp* Rd mutants exhibit a redox potential that is down shifted by about 70 mV,[29,31] mutating from Val to Gly. Interestingly, our computed ISC redox potential +57 mV is close to +39 mV measured for a *Cp* Rd mutant[29] whose sequence [8(G)10(G)11(Y)41(L)43(G)44(G)] involved in ISC H-bonds is closest to model **4**.

We also computed the redox potential of model **4** in AN obtaining −100 mV, which can be compared with the redox potential −136 mV[24] measured in AN for a synthetic peptide-based ISC model with the sequence [8(L)10(G)11(Fp)-41(G)43(G)44(Fp)] (Fp is a Phe-like compound[24]) similar in chemical composition to model **4**. Figure 6 shows the dependence of the computed redox potential of ISC model **4** for different dielectric constants together with the range of measured Rd redox potentials, the range of dielectric constants used to describe an implicit protein model, and the computed and measured redox potentials of the above synthetic ISC model and model **4** in AN.

Our Rd-like ISC model **4** is at the limit of what we can compute at present to evaluate absolute values of redox potentials for ISC models fully quantum chemically. To study the Rd protein in more detail with quantum chemical computations requires more efficient methods like QM/MM[85,86]
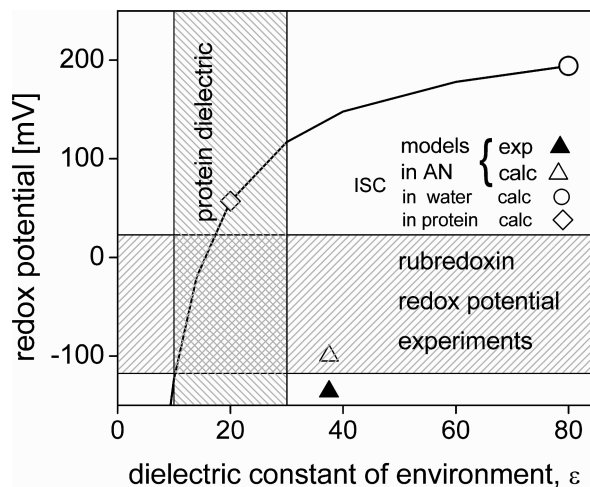
Figure 6. Dependence of the computed redox potential of ISC model **4** as a function of the environmental dielectric constant. The measured Rd redox potentials are within the shaded horizontal regime. The dielectric constants used for implicit protein models are in the shaded vertical regime. The computed redox potentials of ISC model **4** in water and AN are given by the open circle and triangle, respectively. The experimental redox potential for a similar peptide-based ISC in AN is given as the closed triangle.[24] The computed redox potential of model **4** with $\varepsilon = 20$, corresponding to a protein environment is given as the open diamond.

that describe the more distant parts of the complete protein classically and can, therefore, consider also different conformations. However, this approach is beyond the scope of this study. Alternatively, we can use a purely electrostatic approach, which also allows to consider the influence of protonation pattern and hydrogen atom flexibility[87] on the redox potentials of the protein cofactors[78] and the aspects of coupling between electron and proton transfer processess.[88−90] This will be subject of future work.

Factors Shifting Redox Potential in ISC Models. The first factor we discussed was the variation in conformational symmetry between S_4 and C_2. The computed redox potential shift (from model **1a** to **2a**) was 71 mV in water ($\varepsilon = 80$), while for AN ($\varepsilon = 37.5$), the shift was 211 mV. The second factor shifting the ISC redox potential is the number and the type of H-bonds with ISC sulfurs. As we have demonstrated, the shift per H-bond could vary from 65 to 95 mV depending on the type and the geometrical constraints. H-bonds stabilize the negative charge of the ISC in Rd.

The last factor we consider is due to dielectric environment. Comparing at $\varepsilon = 80$, the small ISC model **2a** with the larger ISC model **3h**, both of them without H-bonds, the redox potential decreases from +56 to −238 mV. This demonstrates how a low dielectric can down-shift the ISC redox potential (about 294 mV in this case) when one places a small ISC model in a protein-like environment. A similar down-shift of 221 mV is computed when model **2a** is placed in a low dielectric with $\varepsilon = 20$. The qualitative agreement suggests that the use of $\varepsilon = 20$ for the implicit protein model is appropriate. The results obtained for our ISC models show the importance of the dielectric environment in tuning the

ISC redox potentials and how the degree of solvent exposure of ISC can affect the redox potential.

Inside the protein, these factors may vary in strength and may be coupled. For example, the experiments performed to measure the influence of H-bond strength by means of side chain mutations in rubredoxin mutants[30,33] affect not only the strength of ISC H-bonds but change also the electrostatic volume of the amino acid side chains involved in H-bonds, thus, decreasing the capacity of charge transfer from H-bonds to neighboring atoms.[82,91] Both effects change the redox potential and may act in opposite directions. Hence, in our study, we separated the factors changing ISC redox potentials, making it easier to understand the magnitude and the direction of redox potential shifts depending on the conformations, the H-bonds, and the dielectric environment, all important for ISC proteins.

Conclusions

Using the B4(XQ3)LYP approach, we calculated absolute redox potentials of 17 ISC models related to the ISC in Rd in different media, which are water, AN, and protein-like environment. In AN, the ISC redox potentials are very negative being below −800 mV for the most elementary ISC model with S_4 symmetry, while in water and protein environment, the redox potentials are close to zero. Part of this enormous upshift is due to a change in symmetry from S_4 to C_2, yielding about +200 mV in AN. The remaining part of the redox potential upshift of +600 mV is due to the difference between the aprotic environment provided by AN and the protic environment provided by water. In an implicit solvent model for water, using continuum electrostatics, the energetics of H-bonds are described implicitly by surface charges at the electrostatic boundaries. The ISC models of this study that mimic the protein environment consider polar hydrogen atom contacts to form explicit H-bonds.

The redox potentials computed in the present study agree with available measured values in the range of 50 mV in analogy to the agreement found for the large number of redox potentials, which were computed for transition metal complexes before.[1] This degree of accuracy could only be achieved with the modified B4(XQ3)LYP functional in DFT combined with a post-SCF correction followed by continuum electrostatics to evaluate solvation energies. In this connection, also electrostatic energies were optimized by choosing appropriate atomic radii. Using of the same electrostatics in combination with the B3LYP functional results in an rmsd of 183 mV,[1] comparable with 170 mV reported for calculated redox potentials of 270 different organic compounds.[60]

We explored the influence of two different H-bond partners, Fe−S···H−O and Fe−S···H−N on the ISC redox potentials. We found that in aqueous solution ($\varepsilon = 80$) one Fe−S···H−O H-bond up-shifts the ISC redox potential by 58 mV for the C_2 symmetry and by 90 mV for the S_4 symmetry. In a low dielectric environment ($\varepsilon = 20$) as in proteins, the computed upshift of the ISC redox potential is expected to be 77 mV for a single

Redox Potentials of Rubredoxin Model Complexes

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2907**

Fe−S···H−N H-bond correlating well with the experimental estimate of 70 mV in Rieske proteins.[18] The H-bond influence on ISC redox potentials is larger for Fe−S···H−O than for Fe−S···H−N H-bonds due to the higher polarity of OH groups compared to NH groups in the ISC models of this study. We calculated the redox potential for a model complex that mimics the ISC inside the protein environment and obtained good agreement with experimental values of a Rd-like ISC and an ISC in Rd. This practically quantitative agreement between the computation and the experiment allows to understand the redox potential shift between small ISC models and ISC in proteins, which depends on three factors: the change in conformation, (from symmetry S_4 to C_2, shifting the redox potential by 71 mV), the number of H-bonds (about 60 mV for a single Fe−S···H−N and 90 mV for Fe−S···H−O H-bond) and the degree of solvent exposure (about 200 mV down-shift going from high to low dielectric environment).

Note Added after ASAP Publication. This paper was published ASAP on September 16, 2009. A line describing eq 1 in the Materials and Methods section was corrected. The revised paper was reposted on September 28, 2009.

Supporting Information Available: Listings of Rd pdb information, H-bonds geometrical parameters of ISC models and respective atomic partial charges, charge distribution plot, computed energy tables, and H-bond energies. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Galstyan, A. S.; Knapp, E. W. *J. Comput. Chem.* **2009**, *30*, 203–211.

(2) Hall, D. O.; Cammack, R.; Rao, K. K. *Origins Life Evol. Biosphere* **1974**, *5*, 363–386.

(3) Holm, R. H.; Kennepohl, P.; Solomon, E. I. *Chem. Rev.* **1996**, *96*, 2239–2314.

(4) Beinert, H.; Holm, R. H.; Münck, E. *Science* **1997**, *277*, 653.

(5) Imsande, J. *Plant Physiol. Biochem.* **1999**, *37*, 87–97.

(6) Beinert, H. *J. Biol. Inorg. Chem.* **2000**, *5*, 2–15.

(7) Langen, R.; Jensen, G. M.; Jacob, U.; Stephens, P. J.; Warshel, A. *J. Biol. Chem.* **1992**, *267*, 25625–25627.

(8) Bertini, I.; Ciurli, S.; Luchinat, C. In *Iron-Sulfur Proteins Perovskites*; Springer-Verlag: Berlin, Germany, 1995; Vol. 83, p 1−53.

(9) Stephens, P. J.; Jollie, D. R.; Warshel, A. *Chem. Rev.* **1996**, *96*, 2491–2531.

(10) Eidsness, M. K.; Burden, A. E.; Richie, K. A.; Kurtz, D. M.; Scott, R. A.; Smith, E. T.; Ichiye, T.; Beard, B.; Min, T.; Kang, C. *Biochemistry* **1999**, *38*, 14803.

(11) Maelia, L. E.; Millar, M.; Koch, S. A. *Inorg. Chem.* **1992**, *31*, 4594–4600.

(12) Johnson, R. E.; Papaefthymiou, G. C.; Frankel, R. B.; Holm, R. H. *J. Am. Chem. Soc.* **1983**, *105*, 7280–7287.

(13) Backes, G.; Yoshiki, M.; Loehr, T. M.; Meyer, T. E.; Cusanovich, M. A.; Sweeney, W. V.; Adman, E. T.; Sanders-Loehr, J. *J. Am. Chem. Soc.* **1991**, *113*, 2055–2064.

(14) Zheng, H.; Kellog, S. J.; Erickson, A. E.; Dubauskie, N. A.; Smith, E. T. *J. Biol. Inorg. Chem.* **2003**, *8*, 12–18.

(15) Dey, A.; Jenney, F. E., Jr.; Adams, M. W. W.; Babini, E.; Takahashi, Y.; Fukuyama, K.; Hodgson, K. O.; Hedman, B.; Solomon, E. I. *Science* **2007**, *318*, 1464.

(16) Adman, E. T.; Watenpaugh, K. D.; Jensen, L. H. *P. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 4854–4858.

(17) Rose, K.; Shadle, S. E.; Eidsness, M. K.; Jr., D. M. K.; Scott, R. A.; Hedman, B.; Hodgson, K. O.; Solomon, E. I. *J. Am. Chem. Soc.* **1998**, *120*, 10743–10747.

(18) Hunsicker-Wang, L. M.; Heine, A.; Chen, Y.; Luna, E. P.; Todaro, T.; Fee, J. A. *Biochemistry* **2003**, *42*, 7303–7317.

(19) Solomon, E. I.; Gorelsky, S. I.; Dey, A. *J. Comput. Chem.* **2006**, *27*, 1415–1428.

(20) Carter, C. W. J. *J. Biol. Chem.* **1977**, *252*, 7802–7811.

(21) Low, D. W.; Hill, M. G. *J. Am. Chem. Soc.* **2000**, *122*, 11039–11040.

(22) Nakata, M.; Ueyama, N.; Fuji, M.-A.; Nakamura, A.; Wada, K.; Matsubara, H. *Biochim. Biophys. Acta* **1984**, *788*, 306–312.

(23) Ueyama, N.; Nakata, M.; Fuji, M. A.; Terakawa, T.; Nakamura, A. *Inorg. Chem.* **1985**, *24*, 2190–2196.

(24) Sun, Y.-W.; Ueyama, N.; Nakamura, A. *Inorg. Chem.* **1991**, *30*, 4026–4031.

(25) Walters, M. A.; Dewan, J. C.; Min, C.; Pinto, S. *Inorg. Chem.* **1991**, *30*, 2656–2662.

(26) Huang, J.; Ostrander, R. L.; Rheingold, A. L.; Walters, M. A. *Inorg. Chem.* **1995**, *34*, 1090–1093.

(27) Chung, W. P.; Dewan, J. C.; Turkerman, M.; Walters, M. A. *Inorg. Chim. Acta* **1999**, *291*, 388–394.

(28) Walters, M. A.; Roche, C. L.; Rheingold, A. L.; Kassel, S. W. *Inorg. Chem.* **2005**, *44*, 3777–3779.

(29) Xiao, Z.; Maher, M. J.; Cross, M.; Bond, C. S.; Guss, J. M.; Wedd, A. G. *J. Biol. Inorg. Chem.* **2000**, *5*.

(30) Lin, I. J.; Gebel, E. B.; Machonkin, T. E.; Westler, W. M.; Markley, J. L. *J. Am. Chem. Soc.* **2003**, *125*, 1464–1465.

(31) Park, I. Y.; Eidsness, M. K.; Lin, I. J.; Gebel, E. B.; Youn, B.; Harley, J. L.; Machonkin, E. E.; Frederick, R. O.; Markley, J. L.; Smith, E. T.; Ichiye, T.; Kang, C. *Proteins* **2004**, *57*, 618.

(32) Park, I. Y.; Youn, B.; Harley, J. L.; Eidsness, M. K.; Smith, E.; Ichiye, T.; Kang, C. *J. Biol. Inorg. Chem.* **2004**, *9*, 423.

(33) Lin, I. J.; Gebel, E. B.; Machonkin, T. E.; Westler, W. M.; Markley, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14581–14586.

(34) Bönisch, H.; L., S. C.; P., B.; Ladenstein, R. *J. Biol. Inorg. Chem.* **2007**, *12*, 1163–1171.

(35) Schröter, T.; Hatzfeld, O. M.; Gemeinhardt, S.; Korn, M.; Friedrich, T.; Ludwig, B.; Link, T. A. *Eur. J. Biochem.* **1998**, *255*, 100–106.

(36) Denke, E.; Merbitz-Zahradnik, T.; Hatzfeld, O. M.; Snyder, C. H.; Link, T. A.; Trumpower, B. L. *J. Biol. Chem.* **1998**, *273*, 9085–9093.

(37) Ayhan, M.; Xiao, Z.; Lavery, M. J.; Hamer, A. M.; Nugent, K. W.; Scrofani, S. D. B.; Guss, M.; Wedd, A. G. *Inorg. Chem.* **1996**, *35*, 5902–5911.

(38) Dolan, E. A.; Yelle, R. B.; Beck, B. W.; Fischer, J. T.; Ichiye, T. *Biophys. J.* **2004**, *86*, 2030–2036.

(39) Mouesca, J. M.; Chen, J. L.; Noodleman, L.; Bashford, D.; Case, D. A. *J. Am. Chem. Soc.* **1994**, *116*, 11898–11914.

(40) Li, J.; Nelson, M. R.; Peng, C. Y.; Bashford, D.; Noodleman, L. *J. Phys. Chem. A* **1998**, *102*, 6311–6324.

(41) Ullmann, G. M.; Noodleman, L.; Case, D. A. *J. Biol. Inorg. Chem.* **2002**, *7*, 623–639.

(42) Yelle, R. B.; Park, N.-S.; Ichiye, T. *Proteins* **1995**, *22*, 154–167.

(43) Ergenekan, C. E.; Thomas, D.; Fischer, J. T.; Tan, M.-L.; Eidsness, M. K. *Biophys. J.* **2003**, *85*, 2818–2829.

(44) Tan, M.-L.; Kang, C.; Ichiye, T. *Proteins* **2006**, *62*, 708–714.

(45) Sulpizi, M.; Raugei, S.; Vondele, J. V.; Carloni, P.; Sprik, M. *J. Phys. Chem. B* **2007**, *111*, 3969–3976.

(46) Schmidt am Busch, M.; Knapp, E. W. *J. Am. Chem. Soc.* **2005**, *127* (45), 15730–15737.

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

(48) Noodleman, L.; Norman, J. G., Jr.; Osborne, J. H.; Aizman, A.; Case, D. A. *J. Am. Chem. Soc.* **1985**, *107*, 3418–3426.

(49) Noodleman, L.; Lovell, T.; Liu, T.; Himo, F.; Torres, R. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 259–273.

(50) Torres, R. A.; Lovell, T.; Noodleman, L.; Case, D. A. *J. Am. Chem. Soc.* **2003**, *125*, 1923–1936.

(51) Noodleman, L.; Han, W. G. *J. Biol. Inorg. Chem.* **2006**, *11*, 674–694.

(52) Solomon, E. I.; Xie, X.; Dey, A. *Chem. Soc. Rev.* **2008**, *37*, 623–638.

(53) Dauter, Z.; Wilson, K. S.; Sieker, L. C.; Moulis, J.-M.; Meyer, J. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8836.

(54) Schmidt am Busch, M.; Knapp, E. W. *ChemPhysChem* **2004**, *5*, 1513–1522.

(55) Koch, S. A.; Maelia, L. E.; Millar, M. *J. Am. Chem. Soc.* **1983**, *105*, 5944–5945.

(56) Millar, M.; Lee, J. F.; O'Sullivan, T.; Koch, S. A.; Fikar, R. *Inorg. Chim. Acta* **1996**, *243*, 333–343.

(57) Szilagyi, R. K.; Winslow, M. A. *J. Comput. Chem.* **2006**, *27*, 1385–1397.

(58) Koerner, J. B.; Ichiye, T. *J. Phys. Chem. B* **1997**, *101*, 3633–3643.

(59) Yang, X.; Niu, S.; Ichiye, T.; Wang, L. S. *J. Am. Chem. Soc.* **2004**, *126*, 15790–15794.

(60) Fu, Y.; Liu, L.; Yu, H.-Z.; Wang, Y.-M.; Guo, Q.-X. *J. Am. Chem. Soc.* **2005**, *127*, 7227–7234.

(61) Reiss, H.; Heller, A. *J. Phys. Chem.* **1985**, *89*, 4207–4213.

(62) Trasatti, S. *Pure Appl. Chem.* **1986**, *58*, 955–966.

(63) Lewis, A.; Bumpus, J. A.; Truhlar, D. G.; Cramer, C. J. *J. Chem. Educ.* **2004**, *81*, 596–604.

(64) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998**, *102*, 7787–7794.

(65) Slater, J. C. *Quantum Theory of Molecules and Solids*; McGraw-Hill: New York, 1974.

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

(67) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098–3100.

(68) Lee, C.; Yang, W.; Paar, R. G. *Phys. Rev. B: Solid State* **1988**, *37*, 785–789.

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

(70) Hay, P. J.; Wadt, W. R. *J. Phys. Chem.* **1985**, *82*, 299–310.

(71) *Jaguar, 5.5*; Schroedinger, L.L.C.: Portland, OR, 1991−2003.

(72) Adams, M. W. *Adv. Inorg. Chem.* **1992**, *38*, 341–396.

(73) Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(74) Cornell, W.; Cieplak, P.; Bayly, C.; Kollman, P. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.

(75) Gilson, M.; Honig, B. *Biopolymers* **1986**, *25*, 2097–2119.

(76) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44*, 400–417.

(77) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647–1676.

(78) Ullmann, G. M.; Knapp, E. W. *Eur. Biophys. J.* **1999**, *28*, 533–551.

(79) Bashford, D.; Gerwert, K. *J. Mol. Biol.* **1992**, *224*, 473–486.

(80) Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 9557–9561.

(81) Zhou, P.; Tian, F.; Lv, F.; Shang, Z. *Proteins: Structure, Function, and Bioinformatics* **2009**, *76*, 151–163.

(82) Min, T.; Ergenekan, C. E.; Eidsness, M. K.; Ichiye, T.; Kang, C. *Protein Sci.* **2001**, *10*, 613.

(83) Gillès de Pélichy, L. D.; Smith, E. T. *Biochemistry* **1999**, *38*, 7874–7880.

(84) Day, M. W.; Hsu, B. T.; Joshuator, L.; Park, J. B.; Zhou, Z. H.; Adams, M. W.; Rees, D. C. *Protein Sci.* **1992**, *1*, 1494.

(85) Elstner, M.; Thomas, F.; Suhai, S. *THEOCHEM* **2003**, *632*, 29–41.

(86) Maseras, F.; Morokuma, L. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(87) Kieseritzky, G.; Knapp, E. W. *Proteins: Structure, Function and Bioinformatics* **2008**, *71*, 1335–1348.

(88) Rabenstein, B.; Knapp, E. W. *Biophys. J.* **2001**, *80*, 1141–1150.

(89) Baptista, A. M.; Martel, P. J.; Soares, C. M. *Biophys. J.* **1999**, *76*, 2978–2998.

(90) Onufriev, A.; Case, D. A.; Ullmann, G. M. *Biochemistry* **2001**, *40*, 3413–3419.

(91) Gutman, V.; Resch, G.; Linert, W. *Coord. Chem. Rev.* **1982**, *43*, 133–164.

JCTC Journal of Chemical Theory and Computation

An Evaluation of Explicit Receptor Flexibility in Molecular Docking Using Molecular Dynamics and Torsion Angle Molecular Dynamics

Roger S. Armen,[†] Jianhan Chen,[‡] and Charles L. Brooks III*,[†]

*Department of Chemistry, 930 N. University Ave, University of Michigan,
Ann Arbor, Michigan 48109 and Department of Biochemistry, Kansas State University,
Manhattan, Kansas 66506*

Abstract: Incorporating receptor flexibility into molecular docking should improve results for flexible proteins. However, the incorporation of explicit all-atom flexibility with molecular dynamics for the entire protein chain may also introduce significant error and "noise" that could decrease docking accuracy and deteriorate the ability of a scoring function to rank native-like poses. We address this apparent paradox by comparing the successes of several flexible receptor models in cross-docking and multiple receptor ensemble docking for p38α mitogen-activated protein kinase. Explicit all-atom receptor flexibility has been incorporated into a CHARMM-based molecular docking method using both molecular dynamics (MD) and torsion angle molecular dynamics (TAMD) for the refinement of predicted protein−ligand binding geometries. These flexible receptor models have been evaluated, and the accuracy and efficiency of TAMD sampling is directly compared to MD sampling. Several flexible receptor models are compared, encompassing flexible side chains, flexible loops, multiple flexible backbone segments, and treatment of the entire chain as flexible. We find that, although including side chain and some backbone flexibility is required for improved docking accuracy as expected, docking accuracy also diminishes as additional and unnecessary receptor flexibility is included into the conformational search space. Ensemble docking results demonstrate that including protein flexibility leads to improved agreement with binding data for 227 active compounds. This comparison also demonstrates that a flexible receptor model enriches high-affinity compound identification without significantly increasing the number of false positives from low-affinity compounds.

1. Introduction

It is widely recognized that a significant problem in the field of molecular docking is the incorporation of protein flexibility.[1,2] In experimentally characterized flexible protein binding sites, conformational changes are observed upon binding different ligands that include minor changes of side chain positions, significant local side chain rearrangement, flexible loop rearrangements, larger backbone movements

of the binding pocket, and large hinge motion of domains.[3] Incorporating protein flexibility in molecular docking is challenging because it is difficult to predict the extent of protein conformational changes upon ligand binding (challenges in sampling), and it is also more difficult to score flexible protein−ligand complexes accurately (challenges in scoring). It is clear from the literature that rigid receptor docking can be quite successful for some receptors.[4] However, when rigid receptor docking is used for receptors that are known to be flexible, docking results from a single receptor conformation often suffer from too many false negatives[5] that may be accommodated by minor conformational changes[6] or slightly more open conformations of the

* Corresponding author phone: 734-647-6682; fax: 734-6471604; e-mail: brookscl@umich.edu.

† University of Michigan.

‡ Kansas State University.

binding pocket.[7] Although many studies have shown that incorporating models of protein flexibility can improve docking for these cases, we demonstrate in this article that explicit protein flexibility with all atom molecular dynamics (MD) for the entire protein chain may also introduce significant error and "noise", thereby decreasing docking accuracy and diminishing the ability of a scoring function to discriminate native-like binding poses. In contrast to this result, we also show that similar models that use MD or torsion angle molecular dynamics (TAMD) sampling for flexible side chains and flexible loops are significantly more successful than rigid docking.

In this study, we assess the incorporation of explicit all-atom protein flexibility using both MD and TAMD in CHARMM. Both methods allow flexibility of the protein and ligand simultaneously, which is essential for refinement in protein–ligand conformational space. However, TAMD (internal coordinate space) is more efficient in sampling compared to MD (Cartesian space) because there are significantly fewer degrees of freedom in internal coordinates. The use of conformational dynamics in internal coordinates for flexible receptor docking was pioneered by Abagyan and co-workers.[8,9] Currently, explicitly exploring a large amount of conformational space for both the protein and ligand during the docking procedure is too computationally expensive for screening a large number ($>10^5$) of compounds. As many docking algorithms have been optimized for fast and accurate docking of a flexible ligand into a static protein structure, a computationally attractive approach to incorporate protein flexibility is to use multiple receptor conformations (MRCs) for rigid docking, which is now usually referred to as "receptor ensemble docking" (RED).[10] In this study, we assess several different flexible receptor models using cross-docking accuracy and also compare RED results to experimental binding data.

P38α mitogen-activated protein (MAP) kinase is a difficult test case for cross-docking in that it has a very flexible ligand binding site experimentally characterized by numerous crystal structures. Hydrogen exchange experiments monitored by mass spectrometry have shown that the unphosphorylated inactive state of p38α is significantly more flexible than the doubly phosphorylated active state.[11] The majority of discovered inhibitors successfully bind to this more flexible inactive form of the active site, and some inhibitors bind to an alternative conformation of the binding site known as the "DFG-out" conformation.[12] This alternative conformation of the binding site is denoted the DFG-out conformation because the F169 side chain on the flexible activation loop flips out into the binding site revealing a deep hydrophobic pocket to which ligands are able to bind (Figure 1A). Inhibitors have now been observed binding in the DFG-out conformation in numerous tyrosine kinases and ser/thr kinases.[13] It has been proposed that targeting the enzyme in the flexible inactive (nonphosphorylated) conformation is a good strategy for designing selective kinase inhibitors.[14] Several of these inhibitors have been shown to be selective rather than promiscuous, and this has been attributed to binding interactions in the DFG-out conformation. In this paper, we examine an ensemble of 12 inactive (nonphos-
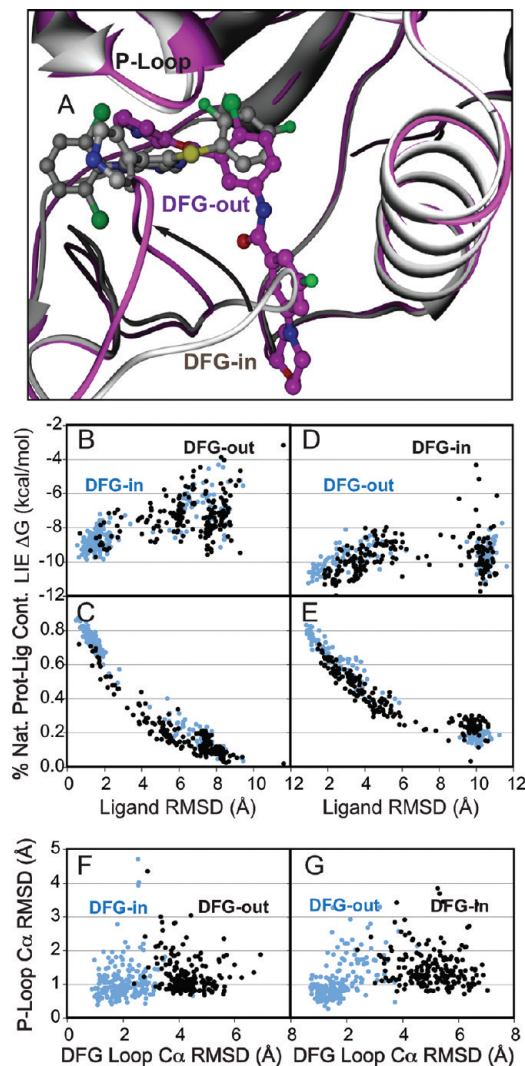

Figure 1. A successful cross-docking for a difficult example involving DFG loop rearrangement. (A) The superposition of a DFG-in protein–ligand complex (1ouy shown in gray) and a DFG-out protein–ligand complex (1w83 shown in magenta). (B) Ligand rmsd versus LIE₁ score plot of flexible receptor self-docking 1ouy (light blue), compared to flexible receptor cross-docking (black) of the ligand 1ouy into the receptor 1w83. (C) Ligand rmsd versus % native protein–ligand contacts for the same. (D) Ligand rmsd vs LIE₁ score plot of flexible receptor self-docking 1w83 (light blue), compared to flexible receptor cross-docking (black) of the ligand 1w83 into the receptor 1ouy. (E) Ligand rmsd versus % native protein–ligand contacts for the same. (F) Cα rmsd of the P-loop versus Cα rmsd of the DFG loop plot for flexible receptor self-docking 1ouy (light blue), compared to flexible receptor cross-docking (black) of the ligand 1ouy into the receptor 1w83. (G) Cα rmsd of the P-loop versus Cα rmsd of the DFG loop plot for flexible receptor self-docking 1w83 (light blue), compared to flexible receptor cross-docking (black) of the ligand 1w83 into the receptor 1ouy.

phorylated) p38α receptor conformations: nine conformationally diverse crystal structures of DFG-in conformations (1a9u, 1bl6, 1bl7, 1di9, 1ouk, 1ouy, 1oz1, 1w84, and 1yqj)[15−20] and three crystal structures of DFG-out conformations (1kv1, 1kv2, and 1w83).[18,21]

2. Results and Discussion

2.1. Rigid Receptor Cross-Docking. With cross-docking simulations, it is possible to determine the sensitivity of docking results to changes in protein conformation by analyzing the effect on the docking accuracy and the prediction of $\Delta G_{binding}$. The entire 12×12 cross-dock was performed from two different starting conformations of the ligand. The first was started from the "swapped conformation" of the ligand from its cocrystal structure into a new crystal structure. The second was started from the "best rotation," which is the lowest energy conformation from 1000 rigid body rotations into the new receptor. Cross-docking results starting from the swapped conformation are biased toward native-like poses and should represent the theoretical maximum cross-docking accuracy for the rigid receptor that atom clashes and conformational changes will allow. Rigid receptor docking results starting from the best rotation are unbiased and more directly analogous to the flexible receptor docking method, which also starts from a best rotation into a flexible receptor conformation. For cross-docking with a rigid receptor, the top five lowest energy docking poses were identified using the R-dependent dielectric (RDIE) scoring function. The reason why it is important to consider the top five scoring poses, and not only the best single scoring pose, is that scoring functions are not always able to correctly rank the most native-like poses, even in the case of self-docking. In comparing cross-docking starting from the swapped conformation and from the best rotation, the results are similar when only the top-scoring pose is considered. However, when the top five scoring poses are examined, it is possible to see that cross-docking from the swapped conformation had a higher percentage (38% ≤ 2.0 Å rmsd) of docking success compared to starting from the best rotation into the new receptor conformation (25% ≤ 2.0 Å rmsd). This is expected, as the conformational search starting from the swapped conformation should be more biased toward low rmsd native-like poses. The overall docking accuracy from rigid receptor cross-docking is similar for the lowest energy pose regardless of whether they are scored using either the RDIE or the LIE_1 (LIE = linear interaction energy) scoring function, but the LIE_1 scores perform slightly better if you consider all of the top five scoring poses. Three receptors (1kv1, 1kv2, and 1w83) are in the DFG-out conformation,[18,21] and when the results from only the single top scoring poses are considered, these receptors are only able to accurately cross-dock their own ligands to within ≤2.0 Å rmsd. Two smaller ligands, 1di9 (a quinazoline-based fragment) and 1w84 (an indole-based fragment), which bind in a similar conformation and are selective to the DFG-in conformation, are also able to bind to the DFG-out receptor conformations within ≤3.0 Å rmsd because the conformation of Phe 169 does not cause a direct atom clash. The ligands selective for the DFG-out conformation are not able to dock accurately in any of the nine DFG-in receptor conformations.

2.2. Constructing Several Models to Represent Protein Flexibility. It was our goal to directly compare cross-docking results using several models of protein flexibility scaling from small side chain movements to large backbone movements.

We first considered a rigid backbone and flexible side chain model and then constructed several flexible protein models that were combinations of flexible side chains and flexible backbone segments. To determine which side chains should be considered flexible, the 12 reference protein−ligand complex structures were superimposed by the best fit to match the backbone C_α coordinates. After superposition, distances less than 6 Å between heavy atoms of the ligand and the protein side chain were used to determine that 34 total residue side chains were considered to interact with the 12 ligands. In the approximation of a rigid backbone, for large residue side chains to adequately sample conformational space, the side chains of adjacent nearest neighbor residues should also be considered to be flexible to allow for concerted structural rearrangements. Therefore, the 34 residues comprising the binding site were linked together into six distinct segments of contiguous residues (res: 30−41, 49−55, 67−89, 103−115, 137−157, 165−177) that form the topology of the binding site. These six contiguous residue segments were then considered to all have flexible side chains, for a total of 74 residues (non-gly or -pro) with flexible side chains. The two most flexible segments in the superposition of the reference structures are residues 30−41, which will be denoted the "P-loop" (phosphate binding loop), and part of the activation loop (res: 165−177), which will be denoted the "DFG loop". These two loops were the most obvious to include as flexible backbone segments in combination with flexible side chains. In total, eight flexible protein models were constructed, which are summarized in Table 1. Model 1 includes flexible side chains only. Models including backbone flexibility were then combined with these flexible side chains: model 2 includes backbone flexibility only in the DFG loop, while model 3 includes backbone flexibility in the two most flexible loops, the P-loop and the DFG loop. Model 4 includes backbone flexibility in all six contiguous residue segments that form the topology of the binding site. Subsequent models 5−7 were constructed and incrementally include flexibility into previously rigid segments. Model 7 is composed of one large flexible segment that defines the entire topology of the binding site (res: 30−177) but still allows 58% of the protein chain to be rigid. Model 8 includes full backbone and side chain flexibility of the entire protein (res: 5−351).

2.3. TAMD Flexible Receptor Cross-Docking. For flexible receptor docking using TAMD, the entire 12×12 cross-dock was performed for the eight different models of protein flexibility. Ligand heavy atom rmsd metrics were used to compare overall docking accuracy over all 144 cross-docking simulations (Table 2). Considering the percentage where the lowest scoring pose is ≤2.0 Å rmsd, the docking accuracy seems to decrease proportionally to the log of the flexible receptor degrees of freedom. The docking accuracies of flexible models 1−6 are all significantly better than rigid docking. Flexible model 8, where the entire protein chain is considered to be flexible, is the only model that had a lower docking accuracy than rigid docking. The fact that the most flexible receptor model, model 8, exhibited the lowest docking accuracy agrees with observations from previous studies, suggesting that increased receptor flexibility may

Table 1. Description of Flexible Receptor Docking Models

name for flex protein model	N degrees of freedom	description	N flex backbone residues	N Flex side chains	residues composing flexible backbone segments
TAMD model 1	271	flex side chains only	0	75	none
TAMD model 2	295	1 flex loop + flex side chains	12	75	DFG loop: 165−177
TAMD model 3	317	2 flex loops + flex side chains	23	75	P-loop: 30−41, DFG loop: 165−177
TAMD model 4	430	6 flex backbone segments	83	75	30−41, 49−55, 67−89, 103−115, 137−157, 165−177
TAMD model 5	514	4 flex backbone seg.	103	N/A	30−41, 49−89, 103−115, 137−177
TAMD model 6	604	2 flex larger backbone seg.	125	N/A	30−115, 137−177
TAMD model 7	720	1 huge flex backbone seg.	147	N/A	30−177
TAMD model 8	1639	entire protein flexible	346	N/A	5−351
MD model 2	2817	1 flex loop + flex side chains	12	75	DFG loop: 165−177
MD model 3	3030	2 flex loops + flex side chains	23	75	P-loop: 30−41, DFG loop: 165−177
MD model 4	4149	6 flex backbone segments	83	75	30−41, 49−55, 67−89, 103−115, 137−157, 165−177
MD model 8	16788	entire protein flexible	346	N/A	5−351

Table 2. Cross-Docking Accuracy for a Rigid Receptor Compared to Eight Flexible Receptor Models

name for protein model	N degrees of freedom	description	best pose ≤2.0 Å (%)	best pose ≤3.0 Å (%)	any of top five ≤2.0 Å (%)	any of top five ≤3.0 Å (%)
rigid receptor 1	0	start with "swapped conf."	15	24	38	51
rigid receptor 2	0	start with "best rotation"	15	24	25	38
rigid receptor 1	0	rescore with flex receptor LIE$_1$	21	31	37	49
TAMD model 1	271	flex side chains only	40	58	61	85
TAMD model 2	295	1 flex loop + flex side chains	37	60	76	97
TAMD model 3	317	2 flex loops + flex side chains	35	51	84	94
TAMD model 4	430	6 flex backbone segments	31	51	60	84
TAMD model 5	514	4 flex backbone seg.	26	44	56	86
TAMD model 6	604	2 flex larger backbone seg.	26	47	58	83
TAMD model 7	720	1 huge flex backbone seg.	20	41	59	84
TAMD model 8	1,639	entire protein flexible	10	30	48	80
MD model 2	2,817	1 flex loop + flex side chains	35	65	65	88
MD model 3	3,030	2 flex loops + flex side chains	33	56	71	91
MD model 4	4,149	6 flex backbone segments	27	50	60	84
MD model 8	16,788	entire protein flexible	12	27	44	82

result in lower overall docking accuracy.[22] However, when you consider the results including the top five scoring poses, model 8 still performs much better than rigid docking. This indicates that model 8 was able to successfully sample and then rank more native-like poses into the top five compared to rigid docking. For model 8, when low rmsd poses are in the top five, there is a diminished ability of the scoring function to correctly rank the most native pose as the best pose compared to the other flexible receptor models. For model 8, when a pose ≤2.0 Å rmsd is represented in the top five, it is correctly ranked as the lowest scoring pose only 21% of the time. In contrast, for model 1, where only the side chains are flexible, when a pose ≤2.0 Å rmsd is represented in the top five, it is correctly ranked as the lowest scoring pose 67% of the time. Models 2−4 are able to correctly rank low rmsd poses from the top five as the top pose on the order of 42−52% of the time. Therefore, the most flexible receptor model displays a reduced ability to score or rank native-like binding poses compared to the other models.

Considering only the percentage where any of the top five scoring poses is ≤2.0 Å rmsd) in Table 2, the docking accuracy increases from rigid receptor (38%) to model 1 (61%), to model 2 (76%) and is the highest for model 3 (84%). Model 3 is the most successfully flexible receptor model for sampling and then ranking low rmsd poses into the top five (Supporting Information Figure S1). From a careful comparison of all 144 cross-docks individually, it is clear that this improvement is due to the inclusion of backbone flexibility for the two loops. The increase in docking accuracy from 38% (rigid receptor) to 61% (model 1) reflects success for many of the less difficult cross-docking situations where only minor or major changes in side chain conformation are sufficient for successful docking. The increase in docking accuracy from 61% for model 1 to 76% for model 2 reflects success in more difficult cross-docking situations where rearrangements in the backbone conformation of the DFG loop and side chains are required for successful docking. Including flexibility in the two most flexible loops (model 3) was even more successful at ranking low rmsd poses (≤2.0 Å) in the top five. It was somewhat surprising that there was such a drop in docking accuracy from model 3 (two flexible loops) to model 4 (six flexible segments). From superposition of the cross-docking crystal structures, it was expected that model 4 would perform well for backbone rearrangements in the gatekeeper region (res: 103−115) required for success in a few specific cross-docks. Even though model 4 outperforms models 1−3 for a few of these specific cross-docking examples, it is less successful in many of the less challenging cross-docks that required only minor changes in side chain conformation.

Our investigation of models 4−8 was aimed at assessing ways of including binding site flexibility in the absence of multiple crystal structures or any evidence as to which parts

of the receptor binding site are flexible. If nothing was known about a receptor's flexibility, would it be better to apply backbone flexibility to multiple protein segments that form the binding site, or to all contiguous residues that form the binding site, or to the entire protein? It was surprising that, when the top five scoring poses were examined, there was not much of a difference between models 4−7. In contrast to models 1−3, models 4−7 include backbone flexibility in the hinge region (res: 106−112), which is the third most flexible backbone segment in the reference crystal structures, and atom clashes involving this region result in some rigid docking failures. It was therefore surprising that including backbone flexibility in this region did not have a significant impact. In general, our results reflect that the best docking success is achieved with the most minimal model that adequately describes the experimentally observed protein flexibility. Model 3, which has many flexible side chains and only two flexible loops, demonstrates the best performance in docking accuracy because it describes the flexibility of the receptor adequately with the fewest necessary degrees of freedom comprising the conformational search space. It would be possible to pursue this strategy for a new receptor target even in the absence of any knowledge of the flexibility of the binding site, by predicting the most flexible segments from MD simulations, normal-mode analysis,[23,24] elastic network normal modes,[25] or other framework approach algorithms.[26] A comparison between flexible receptor models 1−8 also suggests that it would likely be better to consider only the segments forming the binding site to be flexible, rather than the entire protein. These principles can be widely extended to build efficient flexible receptor docking models for new targets of interest. For systems where large conformational changes are expected, our results suggest that it may be a good idea to use other conformational sampling approaches (including normal-mode analysis, fully flexible MD conformational sampling with solvent, etc.) in order to identify a small subset of diverse low-energy conformations. Then, when low energy conformations representing these various states are selected, local explicit flexibility in the binding site can be applied to improve docking accuracy. This approach is in contrast to docking into a large number of fully flexible structures (like flexible model 8) where the performance of a scoring function is diminished due to energetic "noise".

Reduced docking accuracy in flexible model 8 was primarily due to significant binding site rearrangements resulting in low-scoring non-native ligand binding poses. Distortions in the geometry of the ligand binding residue side chains directly resulted in a loss of docking accuracy in flexible model 8. Over the entire cross-dock, we show that 60−70% of the native protein−ligand contacts were sufficient for docking in native-like binding modes. In the majority of the non-native binding pocket conformations, more than 50% of the ligand binding residues had undergone significant rearrangement and cannot form enough native-like ligand contacts. Specific cross-docks, where only model 8 failed, were examined in detail, and in many cases there were significant binding site conformational changes in the side chains that are not observed in any of the reference

crystal structures. In these cases, the overall backbone C_α rmsd is in the range of 1.5−2.0 Å from the starting and reference structures, but these backbone deviations are on the order of the maximum C_α rmsd between reference crystal structures (1.7 Å C_α rmsd for 1w83 and 1yqj). One similarity between these docking failures was that there were significant changes in either the backbone or side chain conformation of the majority of residues that form native contacts with the ligand. Two specific groups of conserved hydrophobic contacts in kinase scaffolds have been proposed as key contacts involved in catalytic and regulatory conformational changes: the regulatory spine (res: 75, 86, 148, and 169) and the catalytic spine (res: 38, 51, 113, 156, 157, 158, 212, and 216).[27,28] An interesting observation was that, in many cases where only model 8 failed, conformational changes were also observed in the side chain contacts in the catalytic spine, where no changes in the catalytic spine are observed between any of the 12 reference crystal structures despite significant changes in the binding pocket. Model 8 is the only model where the entire catalytic spine is flexible, and these observed rearrangements only occurred for model 8. Therefore, it would have been possible to use this information to select low-energy conformations generated from model 8 that maintained native-like structure in the catalytic spine but allowed significant rearrangements of the binding site.

It is possible that some non-native distortions to the structure in fully flexible model 8 may be due to a lack of explicit or implicit solvent during the brief conformational sampling procedure to generate apo receptor conformations. Including implicit solvent during this conformational search and the subsequent refinement of the receptor ligand complexes would be possible, but at a great increase in computational expense. To assess this, we performed a few control simulations of this procedure using the GBMV implicit solvent model and found that the overall level of backbone (C_α rmsd) and side chain rearrangement is very similar. Apo receptor conformations generated with implicit solvent also exhibited similar binding site distortions in side chain conformations that contact the ligand, as well as side chain conformational changes involved in the catalytic spine (most notably residues Leu156, Val158, and Ile212) and in all four residues in the regulatory spine. One notable difference in the binding site conformations generated with implicit solvent is that Asp168 and Glu72 form less close contacts with Lys53, Arg67, and Arg173, presumably because the full charges on these side chains were solvated. For certain specific cross-docking failures, non-native contacts between Asp168 and Lys53 were formed, but there are also examples of close interactions between these residues in some of the reference cocrystal structures (1ouk and 1w84) compared to the others. It remains unclear if the added computational expense of the implicit solvent during receptor conformation generation would actually improve docking accuracy compared to our other models. Performing the entire cross-dock again with this procedure was beyond the scope of this manuscript. A large ensemble of fully flexible protein structures produced with high-quality explicit solvent simulations would likely retain ensemble members that contain such non-native distortions. Docking into this entire ensemble

would still be "noisy", and therefore, we propose that a better strategy would be to select a few low-energy representative states from such an ensemble and then apply local flexibility to the binding site as mentioned above. Our results for this specific kinase system may not be generalized to all other flexible receptors, but our conclusions should hold in the regime where the amount of sampling required for visiting diverse low-energy conformational states (such as DFG-in/DFG-out) is sufficient to rearrange the majority of ligand interacting residues in the binding site.

TAMD flexible receptor docking succeeds in many difficult cases using models 2 and 3 where rigid receptor docking fails. When the cross-docking results are compared to the experimental binding data using the LIE$_2$ scoring functions (Supporting Information Table S1), the flexible receptor results are in much better agreement (1.1 average unsigned error (AUE) kcal/mol), as expected. The poorer agreement with the binding data for the rigid receptor is the result of lower docking accuracy, especially for atom clashes with high-affinity compounds (false negatives). There is a big improvement in docking accuracy for the three receptors in the DFG-out conformation using the TAMD flexible receptor models, where models 2 and 3 perform the best. Receptors in the DFG-out conformation are now able to successfully dock ligands that bind in the DFG-in conformation as well as the converse, ligands selective for DFG-out are also able to successfully dock to DFG-in receptors. A specific example of one of these difficult cross-docks is swapping the ligands from 1ouy (DFG-in) and 1w83 (DFG-out), shown in Figure 1A. The swapped conformation of each ligand has atom clashes with the DFG loop, requiring significant DFG loop conformational changes for successful cross-docking into a native-like conformation. For flexible receptor cross-docking of the ligand 1ouy into the receptor 1w83 (Figure 1B and C), the lowest-energy pose for this cross-dock was 1.5 Å rmsd for the ligand, and the complex contained 71% of the native protein—ligand contacts. For cross-docking of the ligand 1w83 into the receptor 1ouy, the lowest-energy pose was 2.4 Å rmsd for the ligand, and the complex contained 66% of the native protein—ligand contacts. Among the top five nonredundant poses was another pose with 1.8 Å rmsd for the ligand which contained 69% of the native protein—ligand contacts in the complex.

Despite the fact that the DFG loop does not sample the entire range of DFG loop conformations, TAMD models 2 and 3 are still able to successfully cross-dock ligands selective for the DFG-out conformation into receptors in the DFG-in conformation. Upon careful examination of these structures, many individual examples demonstrate that smaller movements of the DFG loop backbone (on the order of 3−6 Å), as well as changes in side chain conformation, are sufficient to remove the primary atom clashes between the native conformation of the ligand and the DFG loop. Thus, for cross-docking these examples, even though the generated conformations of the DFG loop are only able to come as close as 2.5−3.0 Å C$_\alpha$ rmsd to the reference backbone conformation of the DFG-loop (Figure 1E and F), the ligands are still capable of docking in the correct conformation because up to 70% of the native protein—ligand

contacts are properly formed and contribute the strongest anchor points for binding in the correct orientation. TAMD sampling using model 3 was also successful at correctly predicting significant backbone and side chain conformational changes in the P-loop. Figure 2 shows two separate examples of concerted changes in the two flexible loops that lead to successful cross-docking in difficult cases. In both examples in Figure 2, Tyr 35 from the P-loop undergoes a successful rearrangement from the starting conformation (purple) to the final conformation (red), which is in turn closer to the reference target conformation (blue), allowing the ligand (red) to dock in the correct conformation. Arrows are shown to highlight the movement of Tyr 35 from the starting conformation closer to the reference conformation. In the first example (Figure 2A and B), the ligand 1bl7 (DFG-in) is able to dock accurately into the DFG-out conformation because of the backbone and side chain rearrangement of Phe 169. In the second example (Figure 2C and D), the ligand 1yqj (DFG-in) is able to dock accurately into the 1di9 (DFG-in) receptor conformation. Tyr 35 on the P-loop moves over (red) to form a more native-like contact with the ligand, as in the reference structure (blue).

2.4. MD Flexible Receptor Cross-Docking. TAMD sampling was directly compared to MD sampling over the entire cross-dock set for flexible receptor models 2, 3, 4, and 8. Considering only the first docking accuracy metric in Table 2 (the percentage where the lowest scoring pose is ≤2.0 Å rmsd), TAMD docking accuracy is better than MD for models 2−4. MD sampling reduced the docking accuracy for cross-docks requiring significant DFG loop conformational changes. For many difficult cross-docks, TAMD sampling was able to identify a native-like low-energy conformation, and MD sampling was insufficient. For these examples, scatter plots of ligand rmsd versus LIE$_1$ energy show that TAMD sampling produced low-energy native-like conformations, and MD sampling was unable to sample native-like conformations at all (Supporting Information Figure S2). In other cases, MD sampling was sufficient to sample native-like conformations yet displayed a reduced ability to score or rank native-like binding poses compared to TAMD (non-native conformations score better than the native). Despite these deficiencies of MD sampling compared to TAMD sampling, MD model 2 was used for ensemble docking of 227 active compounds to determine if it was sufficient to demonstrate significant improvement over rigid ensemble docking. This choice was made as a compromise between accuracy and speed. While our rigid receptor docking protocol runs within an hour for each ligand to generate the top-ranked poses, fully flexible TAMD model 8 (or MD model 8) has comparative a runtime on the order of 34 h. MD model 2 has a slightly faster runtime on the order of 26 h when using the consfix routine in CHARMM to speed up the nonbonded evaluations. These runtimes are not yet optimized, and all runtimes were evaluated on a single 2.66 GHz Intel Xeon quad-core processor.

2.5. LIE Scoring Functions Used for Receptor Ensemble Docking. In analyzing the results from docking the series of 227 active compounds into the ensemble of 12 receptor conformations, LIE$_1$ (derived from native scoring) was
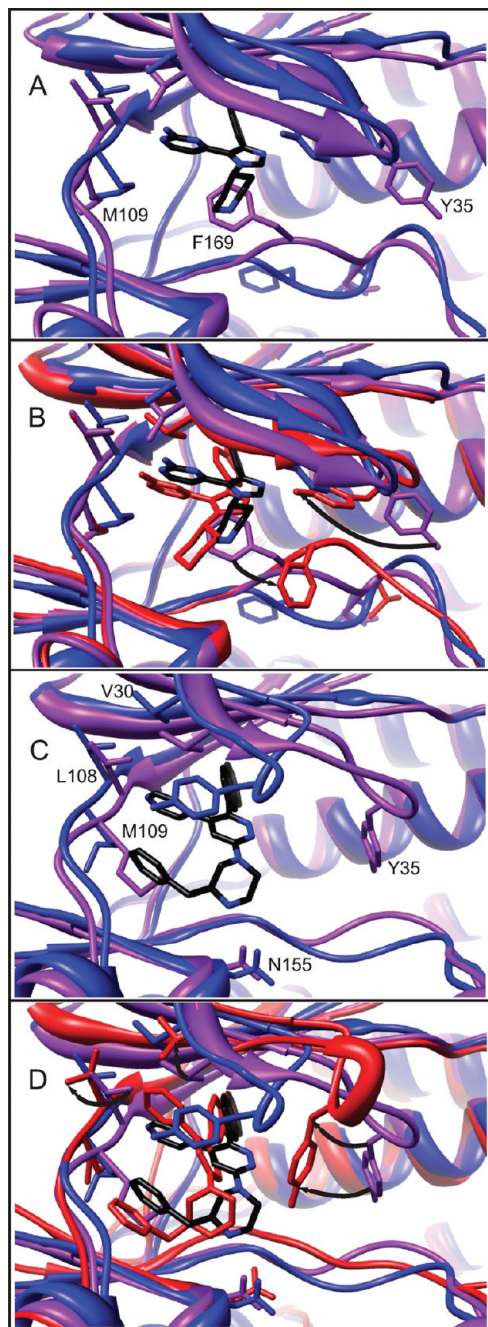
Figure 2. Successful cross-docking examples involving concerted rearrangements of backbone and side chain of the P-loop and the DFG loop. For all images, the reference receptor conformation is shown in blue, superimposed with the starting receptor conformation for a difficult cross-dock shown in purple. The reference ligand position is shown in black. In B and D, the new lowest-energy flexible receptor conformation is shown in red, along with the correctly predicted native-like ligand conformation (rmsd < 2.0 Å). In both examples, Tyr 35 from the P-loop undergoes a successful rearrangement from the starting conformation (purple) to the final conformation (red), which is in turn closer to the reference target conformation (blue), allowing the ligand (red) to dock in the correct conformation. Arrows are shown to highlight the movement of Tyr 35 from the starting conformation closer to the reference conformation. In A and B, the ligand 1bl7 (DFG-in) is able to dock accurately into the DFG-out conformation because of the backbone and side-chain rearrangement of Phe 169. In the second example (C and D), the ligand 1yqj (DFG-in) is able to dock accurately into the 1di9 (DFG-in) receptor conformation.

applied to rank the ligand conformations for both rigid and flexible receptor docking. After the top five scoring poses are identified, calculated energy components from MD are used for comparison to experimental binding data using the LIE_2 model that incorporates desolvation. For clarity in presenting results from ensemble docking, only results from the single lowest-scoring pose will be used, rather than including all results from the top five. Our objective in comparing to binding data (fitting LIE_2 to the experiment) is to assess if the top-scoring ligand poses derived from flexible receptor docking are able to provide better agreement with binding data than rigid receptor docking.

In analyzing results from ensemble docking using the LIE_2 model, much better fits to the experimental data were achieved by adding an entropic term related to the number of freely rotatable bonds[29] in each given ligand that becomes immobilized upon binding to the receptor: $[\Delta G_{binding} = \alpha(\Delta_{VDW}) + \beta(\Delta_{(ELEC+GBE)}) + \gamma(N_{rotatable\_bonds})]$. This term had not been included previously in LIE_2 because the cross-docking test set showed minimal improvement upon adding the entropic term. Fits to the LIE_2 model yield values of γ that are within the range of other values (0.1−0.6 kcal/mol) reported in the literature for the entropic contribution to $\Delta G_{binding}$ from a single freely rotatable bond in a ligand.[29,30] The parameter γ was found to be (rigid $\gamma = 0.19$, flex $\gamma = 0.35$) when the docking results for all 12 receptors were fit simultaneously to binding data. Comparisons to binding data are calculated in two ways: first, all 12 receptors were fit simultaneously to the binding data using a single LIE_2 model, and second, each receptor was fit individually to the experimental data with its own LIE_2 model. The second approach has smaller errors in kilocalories per mole, as expected, but an important finding is that these two fitting approaches both show similar qualitative trends and few differences in comparing rigid and flexible receptors.

2.6. Docking Accuracy for RED. Although there are not crystal structures for all of the 227 active compounds, a reasonable approximation can be made for the ligand binding conformations assuming that the vast majority of active compounds will adopt binding modes similar to known crystal structures of similar compounds. There are six major series of related compounds: (1) 2-aminopyrimidine carbamates,[31] (2) 4-azaindoles,[19] (3) aminopyrazoles,[32] (4) benzimidazolones,[33] (5) indole and pyridine fragments,[18] and (6) triazolopyridine oxazoles.[34] All ligands that bind in the DFG-in conformation share at least two anchor points or pharmacophores (Figure 3): (1) a hydrogen bond or close interactions with the backbone and side chains in the gatekeeper region of the active site (res: 107−110H) and (2) a deep hydrophobic pocket that binds aromatic groups (T106, L75, and L104). These two anchor points can be used to determine the overall orientation of each ligand docking pose. Series 2, 4, and 6 also have a third unique anchor point which is a hydrophilic interaction with the charged NH_3 group of K53. Ligand binding poses were considered to bind in a native-like orientation if they satisfied two or more of these anchor points, and this criterion was used to approximate the overall docking accuracy for each receptor (Table 3). Hydrogen bonding to the hinge region in Figure
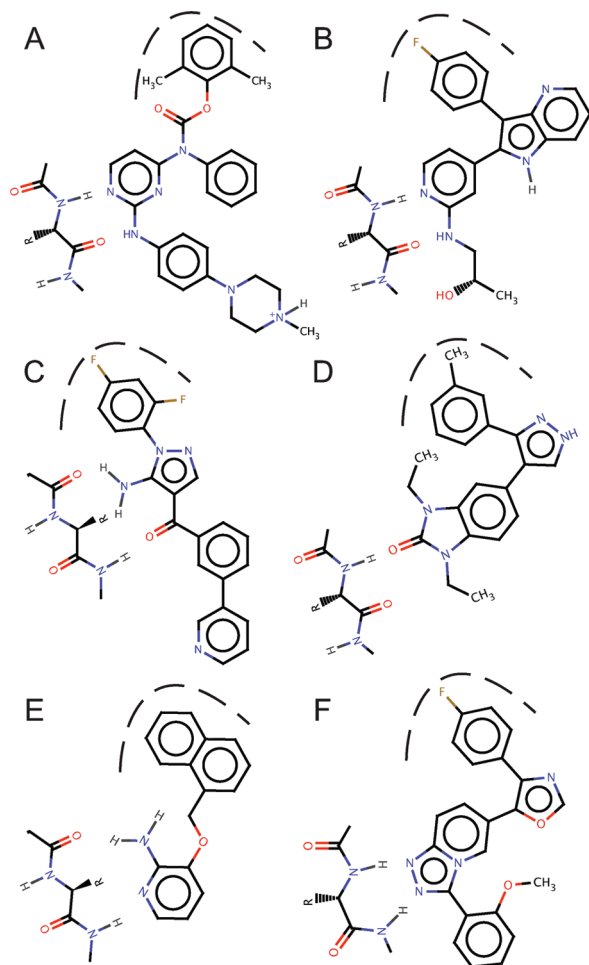
Figure 3. Two common protein–ligand interactions (anchor points), shared by all six series of active compounds binding in the DFG-in receptor conformation. An interaction with the hydrophobic pocket (dashed line) and hydrogen bonding to residues in the gatekeeper region are common to all six series of active compounds: (A) 2-aminopyrimidine carbamates, (B) 4-azaindoles, (C) aminopyrazoles, (D) benzimidazolones, (E) indole and pyridine fragments, and (F) triazolopyridine oxazoles.

3F shows the peptide plane flip (from structure 1zzl) that is also present in the structure for 1ouk and 1ouy. All compounds are able to hydrogen-bond to the hinge region in either conformation, although the ideal ligand interactions may not be formed.

The highest docking accuracy using a rigid receptor model was found for receptors 1ouk (67%) and 1ouy (64%). The receptor 1ouk also had the best docking accuracy in the cross-docking study according to ligand rmsd metrics. This is an interesting finding in light of a recent solution NMR study which concluded that the predominant ensemble in solution is most similar to the crystal structures of 1ouk and 1ouy.[35] This NMR study compared a set of 73 measured residual dipolar couplings (RDCs) from the apo conformation in solution to various crystal structure conformations.[35] Receptor conformations 1ouk and 1ouy exhibit a peptide plane flip over the gatekeeper residues M109 and G111, which participate in backbone hydrogen bonding with ligands. Comparison of the measured RDCs upon binding the pyridinylimidazole inhibitor 1a9u demonstrated evidence for

a structural transition over residues M109 and G111 from the receptor conformation (1ouk) to the crystallographic conformation (1a9u).[35] Therefore, prior to binding, the predominant conformation of the apo receptor was most similar to 1ouk and 1ouy. This experimental observation rationalizes the observed high docking accuracy and good agreement with binding data for rigid receptor docking with 1ouk and 1ouy. It is also interesting that the lowest rigid receptor docking accuracy is observed for the receptor conformation of 1a9u, and this can be attributed to the peptide plane flip in the gatekeeper region as well as a very closed conformation of the P-loop. Low docking accuracy for the ligand 1a9u in cross-docking can also be explained by NMR evidence that this ligand binds in more than one conformation.[35,36]

There are several receptors that perform very poorly for rigid docking that show improvements in docking accuracy upon adding flexibility. The docking accuracy for individual rigid receptors ranges from as low as 11% for 1a9u to as high as 67% for 1ouk. For 5 of 12 receptors with relatively low rigid docking accuracy (1a9u, 1bl6, 1bl7, 1w83, and 1w84), the flexible docking model improves the docking accuracy. However, for five other receptors that had a comparatively high rigid receptor docking accuracy (1di9, 1ouk, 1ouy, 1oz1, and 1yqj), including flexibility slightly diminished overall docking accuracy and overall agreement with binding data. In this way, flexibility improved agreement with experimental results for receptors that perform poorly in rigid docking and reduced agreement with experimental results for the receptors that had the best results for rigid docking. However, despite this small decrease in overall docking accuracy for these receptors, flexible receptor docking still achieved overall improvement in ranking compounds when compared to binding data.

2.7. RED Comparison to Binding Data. Comparison to experimental binding data shows that the flexible receptor model has an improved ability to rank compounds. When all 12 receptors are fit to binding data simultaneously, 8 of 12 have a better linear correlation (R^2) with the experimental data with a flexible receptor. When the receptors are fit individually, 7 of the 12 receptors showed improved agreement in terms of error (AUE) in kilocalories per mole and linear correlation (R^2) with a flexible receptor. In a comparison of each individual receptor to experimental binding data, the flexible receptor model demonstrated improved enrichment of high-affinity compounds over low-affinity compounds (Table 3, Supporting Information Table S2). Results from rigid and flexible docking are compared by calculating the enrichment factor (EF) for each receptor on the basis of the number of high affinity hits in the top n ranked compounds. EFs were calculated [EF = (Hits$_n$/NC$_n$)/(Hits$_{total}$/NC$_{total}$)], where Hits$_n$ is the number of true high-affinity actives retrieved in NC$_n$ (the top n ranked compounds), Hits$_{total}$ is the total number of high-affinity actives, and NC$_{total}$ is the total number of compounds. Enrichment factors were calculated for two sets of the highest-affinity compounds. The first group of high-affinity compounds was 46 of 227 ligands that have $\Delta G_{exp} < -11.0$ kcal/mol, and

Table 3. Receptor Ensemble Docking Results: Docking Accuracy, Comparison to Binding Data, and Enrichment Factor (EF$_1$Max)[a]

id	PDB		docking accuracy	fit all simultaneously				fit each individually			
				AUE (kcal/mol)	rms (kcal/mol)	R^2	EF$_1$ Max	AUE (kcal/mol)	rms (kcal/mol)	R^2	EF$_1$ Max
1	1a9u	rigid	0.11	1.90	2.24	0.04	0.71	1.82	2.16	0.03	0.22
1	1a9u	flex	0.22	2.11	2.54	0.12	1.76	1.68	2.11	0.10	1.57
2	1bl6	rigid	0.20	1.87	2.26	0.05	0.73	1.76	2.16	0.04	0.65
2	1bl6	flex	0.41	1.86	2.23	0.14	1.71	1.65	2.07	0.14	1.62
3	1bl7	rigid	0.32	1.88	2.24	0.04	0.58	1.87	2.23	0.04	0.60
3	1bl7	flex	0.38	1.87	2.29	0.10	0.73	1.67	2.09	0.09	0.68
4	1di9	rigid	0.27	1.78	2.16	0.19	2.42	1.43	1.80	0.18	2.25
4	1di9	flex	0.20	2.03	2.50	0.13	1.47	1.64	2.09	0.12	1.46
5	1kv1	rigid	0.15	2.37	3.13	0.01	0.55	2.32	2.94	0.01	0.52
5	1kv1	flex	0.04	2.83	3.61	0.02	0.83	1.57	1.99	0.12	1.33
6	1kv2	rigid	0.26	2.25	2.90	0.00	0.36	2.23	2.71	0.01	0.33
6	1kv2	flex	0.03	2.05	2.63	0.09	0.76	1.40	1.75	0.22	1.69
7	1ouk	rigid	0.67	1.53	1.85	0.15	0.93	1.49	1.84	0.16	1.11
7	1ouk	flex	0.33	1.95	2.33	0.10	1.22	1.64	2.05	0.10	1.19
8	1ouy	rigid	0.64	1.57	1.92	0.12	0.93	1.57	1.92	0.12	0.89
8	1ouy	flex	0.46	1.97	2.40	0.07	1.31	1.75	2.20	0.06	1.22
9	1oz1	rigid	0.49	1.49	1.87	0.21	1.38	1.41	1.78	0.19	1.07
9	1oz1	flex	0.18	1.79	2.21	0.11	1.16	1.61	2.06	0.12	1.33
10	1w83	rigid	0.05	1.95	2.52	0.00	0.35	1.67	2.11	0.07	1.33
10	1w83	flex	0.08	1.85	2.43	0.06	0.73	1.52	1.93	0.14	1.67
11	1w84	rigid	0.20	1.65	2.08	0.18	1.46	1.53	1.90	0.14	0.95
11	1w84	flex	0.21	1.51	1.89	0.19	2.17	1.49	1.88	0.21	2.50
12	1yqj	rigid	0.50	1.62	1.96	0.08	0.49	1.49	1.85	0.17	0.93
12	1yqj	flex	0.48	1.74	2.11	0.11	1.21	1.60	1.98	0.12	1.30

[a] The enrichment factor EF$_1$ was calculated for the first group ($\Delta G_{exp} < -11.0$ kcal/mol) of high-affinity compounds. The maximum observed EF$_1$ (EF$_1$ Max) within the top 20% of ranked compounds is reported here. The maximum theoretical value of EF$_1$ is 4.9. Additional enrichment factors analysis is reported in Supporting Information Table S2.

the second group was 89 of 227 ligands with $\Delta G_{exp} < -10.0$ kcal/mol. The enrichment factor EF$_1$ was calculated for this first group 1 ($\Delta G_{exp} < -11.0$), and the enrichment factor EF$_2$ was calculated for group 2 ($\Delta G_{exp} < -10.0$ kcal/mol). Given these definitions, the maximum theoretical value of EF$_1$ is 4.9, and the maximum theoretical value of EF$_2$ is 2.6. A third group of 57 low-affinity compounds was selected from the 227 using the criterion $\Delta G_{exp} > -8.0$ kcal/mol. This third group of low-affinity compounds was used to calculate the number of false positives, assuming that group 2 (true positives) could be successfully separated from group 3. A large majority of the individual receptors (10 of 12 receptors) had improved enrichment factors with a flexible receptor (Figure 4), and the overall enrichment over the ensemble of 12 receptors is better for the flexible receptor model (Supporting Information Tables S2 and S3). The flexible receptor model also showed significant enrichment compared to the rigid receptor model in separating these 227 active compounds from a large number (>9000) of decoy molecules (Figure 5A and B). These decoy molecules, which are specific for p38α, were taken from the database of useful decoys (DUD) and represent a very difficult decoy set because the ligands were selected to be similar to the actives in several molecular properties.[37] As 28 of our 227 active compounds are greater in molecular weight (MW) than the DUD decoy molecules, we also calculated the enrichment factors considering only 153 active ligands that share the same molecular weight range (450 > MW > 320) as the DUD decoys to eliminate any source of artificial enrichment (Figure 5C and D).[38] Either approach to calculating the enrichment factor shows significant improvement of the

flexible receptor model compared to the rigid receptor for either DFG-in or DFG-out conformations (Figure 5).

As our results show that certain receptors perform well for one ligand series, while not so well for others, we investigated several combinations of minimal receptor ensembles. A minimal receptor ensemble of four receptors (1a9u, 1di9, 1ouk, and 1oz1) was constructed using the best rigid receptor for each series respectively. When these rigid receptor docking results are fit simultaneously to the LIE$_2$ model, it achieves much better agreement with binding data than any individual receptor (Figure 6) and shows that it is possible for this LIE$_2$ model to rank all six series simultaneously using a MRC ensemble approach. For the six major series of related compounds, the receptor with the best agreement was used for each series respectively: (1) 2-aminopyrimidine carbamates (1ouk), (2) 4-azaindoles (1a9u), (3) aminopyrazoles (1a9u), (4) benzimidazolones (1oz1), (5) indole and pyridine fragments (1di9), and (6) triazolopyridine oxazoles (1ouk). The best linear fit for this model through all points (AUE = 1.1 kcal/mol, rms = 1.4 kcal/mol, and $R^2 = 0.41$) is shown in black in Figure 6A. Figure 6 shows that the best fit receptor for each series has a slightly different slope, even when fit simultaneously. The agreement with the binding data for each series is as follows: (1) 2-aminopyrimidine carbamates ($R^2 = 0.30$), (2) 4-azaindoles ($R^2 = 0.24$), (3) aminopyrazoles ($R^2 = 0.19$), (4) benzimidazolones ($R^2 = 0.16$), (5) indole and pyridine fragments ($R^2 = 0.72$), and (6) triazolopyridine oxazoles ($R^2 = 0.17$).

2.8. Comparison to Other Flexible Receptor Docking Studies with p38α. As p38α is considered to be a representative ser/thr kinase with strong induced fit effects, it has been

2918 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*
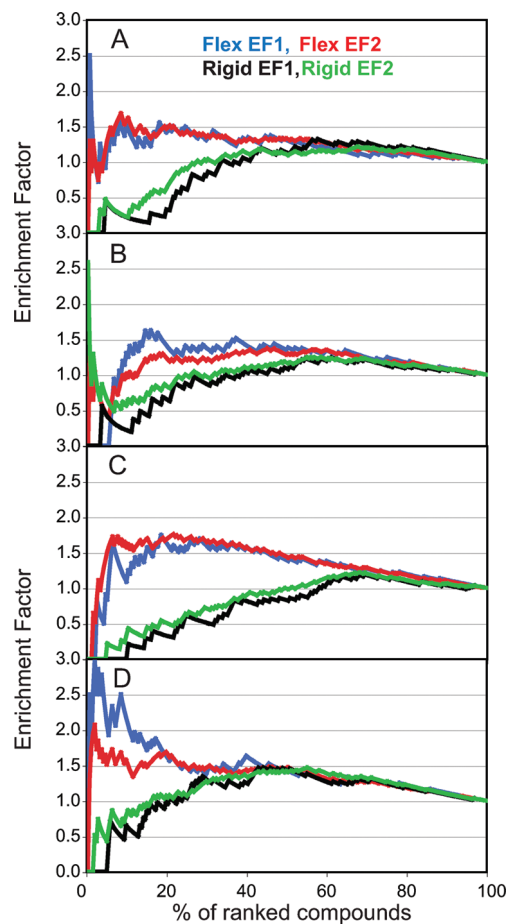
Armen et al.

Figure 4. Enrichment factors for rigid and flexible receptor docking into receptors (A) 1a9u, (B) 1bl6, (C) 1kv2, and (D) 1w84. These enrichment factors EF_1 and EF_2 are calculated assuming that groups of high-affinity compounds can be separated from groups of low-affinity compounds.

used previously in several other flexible receptor docking studies.[39−42] A small subset of our cross-dock can be compared to other studies with ICM[39] and Glide[41] for docking accuracy, where a docking success is defined as the single lowest-energy structure ≤ 2.0 Å rmsd. For a very small 4 × 4 comparison with ICM,[39] our flexible model 1 was the best with a 38% accuracy, which was lower than ICM with a 63% accuracy; however, this lower performance is predominantly from poor docking accuracy of the ligand for 1a9u, which has been shown to bind experimentally in more than one binding mode.[35,36] For a larger comparison with Glide[41] (3 × 3 DFG-out; 9 × 9 DFG-in), flexible model 1 was the best with 46% accuracy, where Glide was 60% accurate.

Early database enrichments were also measured in these two studies using ICM[39] and Glide,[41] and our early enrichment in the top 1−5% of the ranked database (Figure 5C and D) is roughly comparable given that these two studies have different lists of actives and decoys. These two studies also both used on the order of 1000 decoys, and we have used more than 9000 decoys. In the ICM study for the receptor 1a9u, enrichment factors for the top 1%, 2%, and 10% were 16.7, 8.3, and 5.0, respectivly, whereas our corresponding EFs from Figure 5C for 1a9u are 7.1, 3.0, and 1.1, respectively. The early enrichments for our rigid
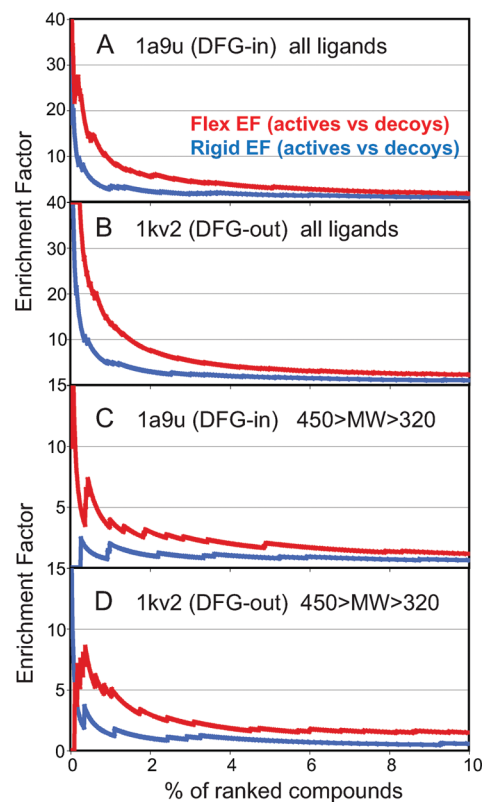

Figure 5. Enrichment factors for separating active compounds from more than 9000 decoy molecules. Enrichment of all 227 actives from decoys for receptor (A) 1a9u and (B) 1kv2. Enrichment of 153 actives in the same molecular weight range (320−450) as the decoys for receptor (C) 1a9u and (D) 1kv2.

docking results for receptor 1kv2 are also reasonable compared to rigid docking performed in the DUD data set paper with DOCK,[37] where the EF at 1% was 2.1 with DOCK and our respective EF at 1% was 1.4, although our EF values range from above 10 to 3.3 for multiple actives found within the top 1%. In a recent report comparing ICM and Glide for enrichment of 111 actives against 6450 decoys, the area under a receiver operating characteristic (ROC) curve was reported for enrichment.[42] In the ROC curve, the signal (fraction of true positives) is plotted on the y axis against the noise (fraction of false positives) on the x axis, and the area under the curve (AUC) indicates the overall signal-to-noise, where 1.0 is theoretically perfect performance and 0.5 is a random performance. In the above-mentioned study, the ROC AUC values for single receptor docking were 0.64 for Glide and 0.70 for ICM, and the ROC AUCs for multiple receptor docking were 0.81 for Glide and 0.88 for ICM.[42]

We have calculated the ROC AUC for our actives with the DUD decoy set (Figure 5C and D) for the molecular weight range of 320−450 and have found that the ROC AUC for the flexible receptor is 0.47 (1a9u) and 0.44 (1kv2). These are both less than 0.5 and are therefore worse than random when considering ranking the entire database, despite the fact that these data sets show reasonable early enrichment within the top 1−10% of the data set. However, the ROC AUCs are still better for the flexible receptor compared to the rigid receptors 0.42 (1a9u) and 0.36 (1kv2). One reason for this poor apparent performance is due to the large number
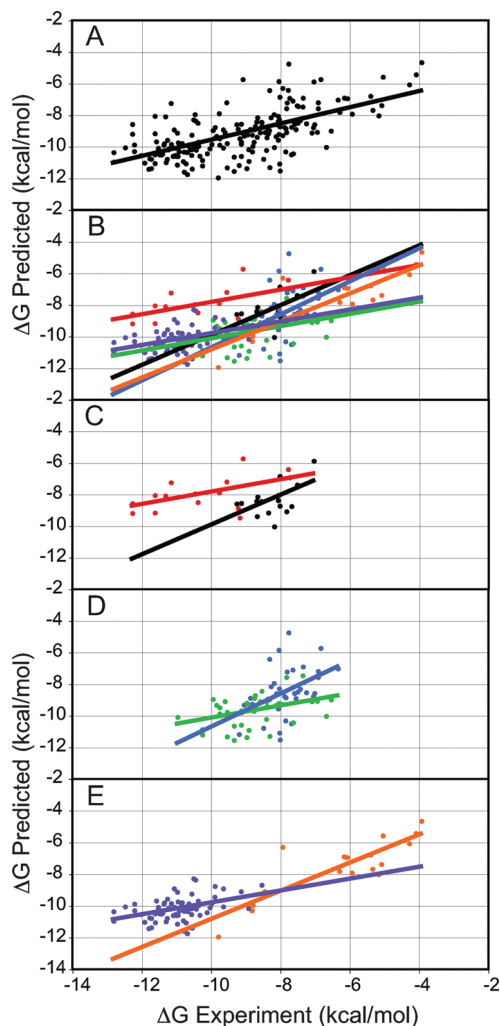
Figure 6. Comparison to binding data, using a minimal ensemble of four rigid receptor conformations (1a9u, 1di9, 1ouk, and 1oz1). The best linear fit for this model through all points is shown in A in black. The best linear fit is shown in B, C, D, and E for each of the six series: (1) 2-aminopyrimidine carbamates (red), (2) 4-azaindoles (black), (3) aminopyrazoles (light blue), (4) benzimidazolones (green), (5) indole and pyridine fragments (orange), and (6) triazolopyridine oxazoles (purple).

of low MW actives in our set, where high MW decoys score better than low MW actives. When only the actives and the decoys in the high MW range of 375−450 are compared (52 actives and 4432 decoys), then our enrichment studies have a ROC AUC of 0.64 (1a9u) and 0.64 (1kv2) for the flexible receptors. These ROC AUC values are both better than random over the entire database and more similar to values reported for single receptor docking for Glide and ICM.[42] Another reason for the lower ROC AUC for the entire MW range of 320−450 may be that the decoys in the DUD data set may be more similar to p38α actives than the decoy set in the study using Glide and ICM.[42] Our results show that, for DUD, reasonable, early enrichment is observed in EF factors but that much better overall enrichment occurs in a comparison of actives and decoys in the MW range of 375−450 compared to the MW range of 320−375. We suspect that this may be because greater MW ligands have

more specific contacts with the receptor compared to their decoys and that lower MW decoys are more similar to low MW actives.[38] In general, although our results for the flexible receptor models are much better than our rigid receptor results, these flexible receptor models are still not performing as well as ICM and Glide.

3. Computational Methods

3.1. Linear Interaction Energy Scoring Function for Docking. For ranking docking poses of the same ligand, using the total energy of the protein−ligand complex with a RDIE is sufficient,[43] as long as the receptor is rigid. When flexibility is introduced into the protein during the conformational search, the total energy of a given protein−ligand complex is no longer a reliable metric to distinguish low-energy native-like poses. This lower predictive power is due to "noise" from conformational changes in intraprotein interactions, which are energetically equivalent or greater than conformational changes in protein−ligand interaction energy. For this reason, when protein flexibility is introduced into the conformational search problem, a scoring function that is based on energetic components from protein−ligand interaction is more effective. In a previous study, CHARMm protein−ligand interaction energies performed well compared to eight other commonly used scoring approaches for the discrimination of docked and misdocked structures.[43] Although this approach works well for rigid docking, for flexible protein docking, we have employed LIE models[44] for our CHARMm protein−ligand interaction energies of the generic form $\Delta G_{\text{binding}} = \alpha(\Delta_{\text{component1}}) + \beta(\Delta_{\text{component2}}) + \gamma(\Delta_{\text{component3}})$ as the sum of potential energy components with weighted coefficients. Following the work of Caflisch and co-workers,[44] we have used a LIE model of the form $\Delta G_{\text{binding}} = \alpha(\Delta_{\text{VDW}}) + \beta(\Delta_{\text{(ELEC+GBE)}})$ for predicting the absolute $\Delta G_{\text{binding}}$ in order to compare ligands of various sizes and shapes (to rank a series of ligands). In the current study of protein flexibility, we use a less computationally expensive LIE model of the form $\Delta G_{\text{binding}} = \alpha(\Delta_{\text{VDW}}) + \beta(\Delta_{\text{ELEC}})$ for ranking binding poses of the same ligand where both the protein and ligand are flexible simultaneously. This simplified LIE score avoids the computationally expensive calculation of generalized Born (using GBMV) solvation energy components for every docking pose. Using this model assumes that differences in solvation energy should be a minor component when considering poses of the same ligand. Therefore, our approach to flexible receptor docking uses a hierarchical two-step scoring approach to reduce computational expense,[45,46] where a simplified "LIE₁" model is used to select the best-scoring poses, which are then rescored using a more sophisticated and computationally expensive "LIE₂" model that includes desolvation.

In order to construct these LIE scoring functions for flexible receptor docking, energetic components acquired from "native scoring" and from self-docking using a rigid receptor were compared. Native scoring LIE components were calculated from both the minimized static native structure and MD simulations started from these structures. Three metrics of comparison are used to compare predicted

$\Delta G_{\text{binding}}$ to experimental $\Delta G_{\text{binding}}$ values: average unsigned error (AUE) in kilocalories per mole, root-mean-square error (rms) in kilocalories per mole, and the best fit linear correlation coefficient (R^2). When the parameters α and β are fit to the energy components from native scoring, the more sophisticated LIE$_2$ model including the desolvation effects [$\Delta G_{\text{binding}} = \alpha(\Delta_{\text{VDW}}) + \beta(\Delta_{\text{(ELEC+GBE)}})$] demonstrates an improved agreement with experimental results as expected compared to LIE$_1$ without solvation [$\Delta G_{\text{binding}} = \alpha(\Delta_{\text{VDW}}) + \beta(\Delta_{\text{ELEC}})$]. The best fit (AUE = 1.0 kcal/mol) was from the LIE$_2$ model applied to ensemble averages over MD simulation. Given the reasonable transferability of these LIE$_1$ fits between native scoring and self-docking, we used them as a scoring function for flexible receptor docking.[44] As an additional check, all of the docking poses collected from self-docking were rescored using the LIE$_1$ scoring function, and native-like ligand poses with heavy atom rmsd ≤ 2.0 Å were still obtained as the lowest-energy poses. Cross-docking results described below show that using LIE$_1$ scores results in improved cross-docking accuracy compared to using the total energy of the complex, for both rigid receptor docking and for flexible receptor docking.

3.2. Preparation of Cross-Docking Set. Our molecular docking approach uses the program CHARMM (Chemistry at HARvard Molecular Mechanics) for an all-atom potential energy description of protein−ligand complexes.[47] The CHARMm force field, originally parametrized by Momany and Rone[48] has been extended to describe ligands in the Ligand−Protein Database (LPDB) and was used to build potential energy functions for all ligands in this study.[49] The crystal structure coordinates of 12 protein−ligand complex structures of p38α MAP kinase were downloaded from the PDB database.[15−21] All structural water molecules were removed, and standard protonation states for neutral pH were applied to all ionizable protein residues. Appropriate protonation states were applied to ligands by calculating pK_a's for titratable groups using both Marvin (http://www.chemaxon.com) and the Web server SPARC (http://ibmlc2/chem.uga.edu/sparc/).[50] Hydrogen atoms were built for ligands using open Babel (http://openbabel.org), and all protein hydrogen atoms were built using CHARMM. For ligands with multiple possible tautomers (such as 1a9u), only the single tautomer that allows a close favorable interaction (between the NH$_3$ group of Lys 53 and imidazole nitrogen) inferred from the cocrystal structure was used.

The activation loop residues D168 and F169 form the characteristic inactive "DFG-out conformation" in the structures of 1kv1 and 1kv2 and 1w83. The structures 1kv1 and 1kv2 are missing electron density for the rest of the flexible residues 170−184 in the activation loop. Flexible loop conformations for this missing electron density were built using Modeler and the MODloop server (http://modbase.compbio.ucsf.edu/modloop/modloop.html).[51] For 1kv1 and 1kv2, several conformations of the activation loop were considered, but the lowest-energy models similar to the DFG-out conformation of 1w83 were selected. The all-atom protein−ligand complexes were minimized for 200 steps in vacuo, and this was taken as the experimental reference conformation of all structures.

3.3. Rigid Receptor Docking (CDOCKER). In our rigid receptor-docking model, a 1.0 Å grid is used to describe the static protein conformation of the binding site, where the interaction energies of 20 types of probe atoms are calculated for every point on the grid. The 3D grid is calculated to extend 8 Å in all directions from any atom in the ligand. Following a simulated annealing conformational search of the flexible ligand, the grid is removed and the all-atom protein−ligand representation is minimized, fixing the coordinates of the protein, using the standard all-atom potential function with a distance-dependent dielectric (RDIE). This interaction energy is taken as the score for the final ligand pose. Other details of previous CDOCKER setups and protocols have been published elsewhere.[52−55] For each given protein−ligand complex, 10 docking trials were performed, composed of 1000 individual docking attempts (100 random conformations × 10 random rotations) for a total of 10 000 docking attempts per complex. Within each docking trial, the interaction energy scores of the final ligand poses are sorted and clustered to identify the "top five" nonredundant clusters of docking poses. After these "top five" docking poses are identified, then a short MD simulation of the entire protein−ligand complex is performed for each of the "top five" in order to more accurately predict $\Delta G_{\text{binding}}$. In these short MD simulations, the electrostatic contribution to the solvation energy is approximated by using a generalized Born implicit solvent potential energy term (GBMV).[56,57] For the "top five" docking poses, MD simulations are performed using a 16 Å nonbonded cutoff and the GBMV implicit solvent model. These MD simulations are performed at 300 K using harmonic restraints on the C$_\alpha$ atoms of the protein backbone, starting with 500 (2 fs) MD steps for equilibration. The harmonic restraints are then released for an additional 1000 (2 fs) steps of production MD, which is used to calculate the ensemble average protein−ligand interaction energies. The first 50 steps are skipped, and then ensemble average interaction energy properties are calculated from 950 structures for the LIE$_2$ model.

3.4. TAMD Flexible Receptor Docking. The TAMD method uses an efficient Newton−Euler inverse mass operator algorithm for solving the equations of motion in internal coordinates.[58] Molecules are represented with a branched tree structure of rigid body clusters connected by torsional hinges. Torsional cross-terms are constructed from local molecular fragments, using a soft-core potential to introduce implicit bond and angle flexibility into the rigid geometry approximation. The use of the soft-core potential effectively removes high-energy barriers on the IC potential energy surface and facilitates more efficient sampling of conformational space. Additional details about the TAMD method have been published elsewhere.[59]

The flexible receptor docking method differs from the rigid docking method in that it never uses a grid representation of the protein. An overview of the flexible docking procedure is shown in Figure 7. In the first step of the flexible receptor docking procedure, random conformations of the ligand are generated for a total of 200 ligand conformations. The next step is to perform TAMD conformational sampling of the
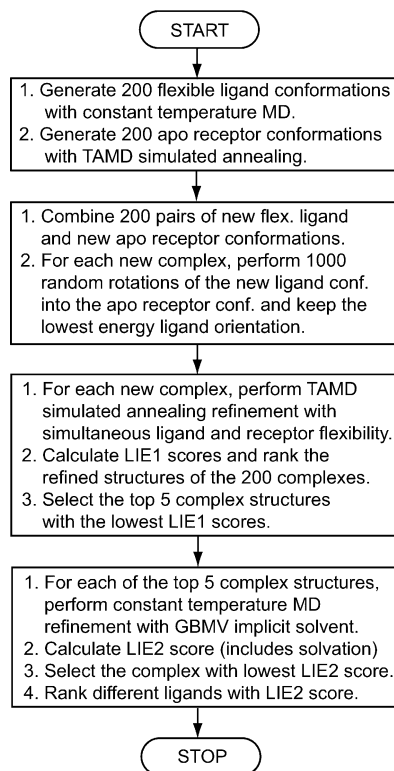
Figure 7. Overview of the flexible receptor docking protocol.

receptor in the absence of the ligand to generate a diverse ensemble of 200 flexible receptor conformations. During this conformation generation procedure, a 14 Å nonbonded cutoff is used along with a RDIE and a soft-core potential. When TAMD is used for simulated annealing, quite large temperatures can be effectively employed for rapid conformational sampling, depending on which model of protein flexibility is used. Models with fewer flexible degrees of freedom can use larger temperatures for sampling without integrator stability problems. For flexible model 2, heating is performed from 2000 to 2500 K, while for flexible model 8, heating is performed from 500 to 700 K and then the system is cooled from 700 K down to 300 K. These protocols yield overall protein rmsd's that are in the range of 0.3–3.0 Å C_α rmsd from the starting receptor conformations.

In the next step, the 200 random flexible receptor conformations are docked with 200 random flexible ligand conformations. For each new random pair of flexible ligand and receptor conformations, 1000 ligand rotations are performed around a reference ligand's center of mass, to identify the optimal rotation conformation of the flexible ligand in the flexible receptor conformation. For this search for an optimal rotation in the binding site, a soft-core potential and all-atom models are used for both the ligand and the receptor. Next, the structure of the protein–ligand complex is refined using TAMD simulated annealing, which allows for simultaneous flexibility of the protein and ligand. The heating and cooling cycle is carried out for all complexes using 1000 steps of 2 fs TAMD, where the system is heated from 500 to 700 K over 1000 steps and then cooled from 700 K back down to 300 K over 1000 steps. During this simulated annealing cycle, significant rearrangements of both

the protein and ligand are possible. Although some complexes do not move far from their starting structure because they contain favorable protein–ligand interactions, other less favorable complexes undergo significant conformational rearrangements. For flexible protein model 2, changes characterized by atom movements as large as 6–8 Å are observed for both ligand and flexible protein side chains in numerous examples. The final TAMD generated protein–ligand complex structures are then minimized for 1000 steps using the soft-core potential. Then, the soft-core potential is removed and replaced with the standard potential function with a distance-dependent dielectric (RDIE). Interaction energies and potential energy components are calculated from this final minimized protein–ligand complex and are used to calculate the LIE_1 score to energy-rank the TAMD generated flexible receptor complexes. Similar to rigid receptor docking, these LIE_1 scores for the final flexible protein–ligand complex conformation are sorted and clustered to identify the "top five" nonredundant clusters of docking poses. As for rigid receptor docking, then a short MD simulation of the entire protein–ligand complex is performed using the GBMV implicit solvent model for each of the "top five" in order to calculate the potential energy components for the LIE_2 score.

3.5. RED for a Series of 227 Active Compounds. In order to extend our study from the $N = 12$ cross-dock, we also considered the docking of 227 active compounds that have IC_{50} binding data for human p38α reported in the Binding Database (www.bindingdb.org).[60] Approximate binding free energies were calculated from IC_{50} data using solution information and the relationship ($K_i = IC_{50}(K_m/(S + K_m))$). The covalent structure and initial low-energy 3D conformations of each ligand were generated from smiles strings using Marvin. Appropriate protonation states were applied to ligands by calculating pK_a's for potential titratable groups using both Marvin and the Web server SPARC. Decoy molecules specific for p38α were taken from the DUD, and the protonation states from DUD were used.[37] All 227 ligands were docked to the ensemble of 12 receptor conformations using both the rigid and flexible receptor-docking model. For flexible receptor docking, flexible protein model 2 was used with MD flexible receptor conformational sampling in order to determine if MD sampling alone was sufficient to achieve improvements over rigid docking.

4. Conclusions

Several models were constructed to represent the flexibility of the binding site for p38α MAP kinase, which exhibits significant binding site flexibility in ligand-bound cocrystal structures. These flexible receptor models scaled from the inclusion of (1) flexible side chains (model 1), (2) one or more flexible backbone segments (models 2 through 7), and (3) treatment of the entire protein as flexible (model 8). As constructed, these models scale in the number of total protein degrees of freedom that describe the conformational search space. Models 2 and 3 demonstrate superior performance in cross-docking accuracy because they describe the flexibility of the system adequately with the fewest degrees of freedom. We show that the fully flexible protein model 8 suffers from

2922 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Armen et al.

reduced docking accuracy. In fully flexible protein model 8, the scoring function docking accuracy is diminished and more "noisy" due to extensive binding site rearrangements involving the majority of the ligand interacting residues. While some of these conformational changes are native and observed in crystal structures leading to docking successes, other non-native rearrangements result in low-scoring non-native ligand binding poses.

In a direct comparison of TAMD sampling to MD sampling for these various models of protein flexibility, TAMD sampling was found to be both more efficient and more accurate. However, even using MD sampling, the percentage of cross-docking success is significantly greater with the flexible receptor models 2 and 3 than for the rigid receptor models. This comparison demonstrates that the flexible receptor models succeed in many difficult cases where the rigid receptor model fails and that reasonable "native-like" receptor conformations are sampled regardless of the starting conformation. Using MD sampling and flexible receptor model 2, we performed RED using the same 12 receptor conformations and 227 active compounds. The comparison to experimental binding data showed that the flexible receptor model showed improved agreement with binding data and an improved ability to rank compounds. In a comparison of each individual receptor, the flexible receptor model also showed improved enrichment of true positives in top-ranked compounds. The flexible receptor methodology described in this paper is most applicable for proteins that show strong induced fit effects.

Abbreviations

fs, femtosecond; ps, picosecond; MD, molecular dynamics; TAMD, torsion angle molecular dynamics; LPDB, Ligand−Protein Database; RDIE, distance-dependent dielectric function; LIE, linear interaction energy; ROC, receiver operating characteristic; AUC, area under the curve; MW, molecular weight; AUE, average unsigned error; rmsd, root mean squared deviation.

Supporting Information Available: Supporting Information is available in the form of two additional figures (S1 and S2) and three tables (S1, S2, and S3). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Carlson, H. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–52.

(2) Abagyan, R.; Totrov, M. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375–82.

(3) Gerstein, M.; Krebs, W. *Nucleic Acids Res.* **1998**, *26*, 4280–90.

(4) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. *Acc. Chem. Res.* **1994**, *27*, 117–123.

(5) Polgar, T.; Baki, A.; Szendrei, G. I.; Keseru, G. M. *J. Med. Chem.* **2005**, *48*, 7946–59.

(6) Zavodszky, M. I.; Kuhn, L. A. *Protein Sci.* **2005**, *14*, 1104–14.

(7) McGovern, S. L.; Shoichet, B. K. *J. Med. Chem.* **2003**, *46*, 2895–907.

(8) Mazur, A. K.; Dorofeev, V. E.; Abagyan, R. A. *J. Comput. Phys.* **1991**, *92*, 261–272.

(9) Totrov, M.; Abagyan, R. *Proteins: Struct., Funct., Genet.* **1997**, 215–220.

(10) Totrov, M.; Abagyan, R. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–84.

(11) Sours, K. M.; Kwok, S. C.; Rachidi, T.; Lee, T.; Ring, A.; Hoofnagle, A. N.; Resing, K. A.; Ahn, N. G. *J. Mol. Biol.* **2008**, *379*, 1075–93.

(12) Akella, R.; Moon, T. M.; Goldsmith, E. J. *Biochim. Biophys. Acta* **2008**, *1784*, 48–55.

(13) Ghose, A. K.; Herbertz, T.; Pippin, D. A.; Salvino, J. M.; Mallamo, J. P. *J. Med. Chem.* **2008**, *51*, 5149–71.

(14) Liu, Y.; Gray, N. S. *Nat. Chem. Biol.* **2006**, *2*, 358–64.

(15) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. *Structure* **1998**, *6*, 1117–28.

(16) Shewchuk, L.; Hassell, A.; Wisely, B.; Rocque, W.; Holmes, W.; Veal, J.; Kuyper, L. F. *J. Med. Chem.* **2000**, *43*, 133–8.

(17) Fitzgerald, C. E.; Patel, S. B.; Becker, J. W.; Cameron, P. M.; Zaller, D.; Pikounis, V. B.; O'Keefe, S. J.; Scapin, G. *Nat. Struct. Biol.* **2003**, *10*, 764–9.

(18) Gill, A. L.; Frederickson, M.; Cleasby, A.; Woodhead, S. J.; Carr, M. G.; Woodhead, A. J.; Walker, M. T.; Congreve, M. S.; Devine, L. A.; Tisi, D.; O'Reilly, M.; Seavers, L. C.; Davis, D. J.; Curry, J.; Anthony, R.; Padova, A.; Murray, C. W.; Carr, R. A.; Jhoti, H. *J. Med. Chem.* **2005**, *48*, 414–26.

(19) Trejo, A.; Arzeno, H.; Browner, M.; Chanda, S.; Cheng, S.; Comer, D. D.; Dalrymple, S. A.; Dunten, P.; Lafargue, J.; Lovejoy, B.; Freire-Moar, J.; Lim, J.; McIntosh, J.; Miller, J.; Papp, E.; Reuter, D.; Roberts, R.; Sanpablo, F.; Saunders, J.; Song, K.; Villasenor, A.; Warren, S. D.; Welch, M.; Weller, P.; Whiteley, P. E.; Zeng, L.; Goldstein, D. M. *J. Med. Chem.* **2003**, *46*, 4702–13.

(20) Tamayo, N.; Liao, L.; Goldberg, M.; Powers, D.; Tudor, Y. Y.; Yu, V.; Wong, L. M.; Henkle, B.; Middleton, S.; Syed, R.; Harvey, T.; Jang, G.; Hungate, R.; Dominguez, C. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2409–13.

(21) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. *Nat. Struct. Biol.* **2002**, *9*, 268–72.

(22) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. *J. Med. Chem.* **2004**, *47*, 45–55.

(23) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. *J. Am. Chem. Soc.* **2005**, *127*, 9632–40.

(24) May, A.; Zacharias, M. *Proteins* **2008**, *70*, 794–809.

(25) Tama, F.; Wriggers, W.; Brooks, C. L., III. *J. Mol. Biol.* **2002**, *321*, 297–305.

(26) Gohlke, H.; Kuhn, L. A.; Case, D. A. *Proteins* **2004**, *56*, 322–37.

(27) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Eyck, L. F. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17783–8.

(28) Kornev, A. P.; Taylor, S. S.; Ten Eyck, L. F. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14377–82.

(29) Salaniwal, S.; Manas, E. S.; Alvarez, J. C.; Unwalla, R. J. *Proteins* **2007**, *66*, 422–35.

(30) Gohlke, H.; Klebe, G. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2644–76.

(31) Martin, M. W.; Newcomb, J.; Nunes, J. J.; McGowan, D. C.; Armistead, D. M.; Boucher, C.; Buchanan, J. L.; Buckner, W.; Chai, L.; Elbaum, D.; Epstein, L. F.; Faust, T.; Flynn, S.; Gallant, P.; Gore, A.; Gu, Y.; Hsieh, F.; Huang, X.; Lee, J. H.; Metz, D.; Middleton, S.; Mohn, D.; Morgenstern, K.; Morrison, M. J.; Novak, P. M.; Oliveira-dos-Santos, A.; Powers, D.; Rose, P.; Schneider, S.; Sell, S.; Tudor, Y.; Turci, S. M.; Welcher, A. A.; White, R. D.; Zack, D.; Zhao, H.; Zhu, L.; Zhu, X.; Ghiron, C.; Amouzegh, P.; Ermann, M.; Jenkins, J.; Johnston, D.; Napier, S.; Power, E. *J. Med. Chem.* **2006**, *49*, 4981–91.

(32) Goldstein, D. M.; Alfredson, T.; Bertrand, J.; Browner, M. F.; Clifford, K.; Dalrymple, S. A.; Dunn, J.; Freire-Moar, J.; Harris, S.; Labadie, S. S.; La Fargue, J.; Lapierre, J. M.; Larrabee, S.; Li, F.; Papp, E.; McWeeney, D.; Ramesha, C.; Roberts, R.; Rotstein, D.; San Pablo, B.; Sjogren, E. B.; So, O. Y.; Talamas, F. X.; Tao, W.; Trejo, A.; Villasenor, A.; Welch, M.; Welch, T.; Weller, P.; Whiteley, P. E.; Young, K.; Zipfel, S. *J. Med. Chem.* **2006**, *49*, 1562–75.

(33) Dombroski, M. A.; Letavic, M. A.; McClure, K. F.; Barberia, J. T.; Carty, T. J.; Cortina, S. R.; Csiki, C.; Dipesa, A. J.; Elliott, N. C.; Gabel, C. A.; Jordan, C. K.; Labasi, J. M.; Martin, W. H.; Peese, K. M.; Stock, I. A.; Svensson, L.; Sweeney, F. J.; Yu, C. H. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 919–923.

(34) McClure, K. F.; Letavic, M. A.; Kalgutkar, A. S.; Gabel, C. A.; Audoly, L.; Barberia, J. T.; Braganza, J. F.; Carter, D.; Carty, T. J.; Cortina, S. R.; Dombroski, M. A.; Donahue, K. M.; Elliott, N. C.; Gibbons, C. P.; Jordan, C. K.; Kuperman, A. V.; Labasi, J. M.; Laliberte, R. E.; McCoy, J. M.; Naiman, B. M.; Nelson, K. L.; Nguyen, H. T.; Peese, K. M.; Sweeney, F. J.; Taylor, T. J.; Trebino, C. E.; Abramov, Y. A.; Laird, E. R.; Volberg, W. A.; Zhou, J.; Bach, J.; Lombardo, F. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4339–44.

(35) Honndorf, V. S.; Coudevylle, N.; Laufer, S.; Becker, S.; Griesinger, C. *Angew. Chem., Int. Ed. Engl.* **2008**, *47*, 3548–51.

(36) Vogtherr, M.; Saxena, K.; Hoelder, S.; Grimme, S.; Betz, M.; Schieborr, U.; Pescatore, B.; Robin, M.; Delarbre, L.; Langer, T.; Wendt, K. U.; Schwalbe, H. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 993–7.

(37) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789–801.

(38) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.

(39) Cavasotto, C. N.; Abagyan, R. A. *J. Mol. Biol.* **2004**, *337*, 209–25.

(40) Subramanian, J.; Sharma, S. *ChemMedChem* **2008**, *3*, 336–44.

(41) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–7.

(42) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. *J. Chem. Inf. Model.* **2008**, *48*, 489–97.

(43) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. *J. Med. Chem.* **2004**, *47*, 3032–47.

(44) Kolb, P.; Huang, D.; Dey, F.; Caflisch, A. *J. Med. Chem.* **2008**, *51*, 1179–1188.

(45) Hoffmann, D.; Kramer, B.; Washio, T.; Steinmetzer, T.; Rarey, M.; Lengauer, T. *J. Med. Chem.* **1999**, *42*, 4422–33.

(46) Yoon, S.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.

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

(48) Momany, F. A.; Rone, R. *J. Comput. Chem.* **1992**, *13*, 888–900.

(49) Roche, O.; Kiyama, R.; Brooks, C. L., III. *J. Med. Chem.* **2001**, *44*, 3592–8.

(50) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.

(51) Fiser, A.; Sali, A. *Bioinformatics* **2003**, *19*, 2500–1.

(52) Vieth, M.; Hirst, J. D.; Dominy, B. N.; Daigler, H.; Brooks, C. L. *J. Comput. Chem.* **1998**, *19*, 1623–1631.

(53) Vieth, M.; Hirst, J. D.; Kolinski, A.; Brooks, C. L. *J. Comput. Chem.* **1998**, *19*, 1612–1622.

(54) Wu, G. S.; Robertson, D. H.; Brooks, C. L.; Vieth, M. *J. Comput. Chem.* **2003**, *24*, 1549–1562.

(55) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., III. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755–63.

(56) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–56.

(57) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 265–84.

(58) Jain, A.; Vaidehi, N.; Rodriguez, G. *J. Comput. Phys.* **1993**, *106*, 258–268.

(59) Chen, J.; Im, W.; Brooks, C. L., III. *J. Comput. Chem.* **2005**, *26*, 1565–78.

(60) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucleic Acids Res.* **2007**, *35*, D198–201.

(61) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605–12.