JCTC Journal of Chemical Theory and Computation

Efficient Diffuse Basis Sets for Density Functional Theory

Ewa Papajak and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431

Received October 24, 2009

Abstract: Eliminating all but the *s* and *p* diffuse functions on the non-hydrogenic atoms and all diffuse functions on the hydrogen atoms from the aug-cc-pV(*x*+d)Z basis sets of Dunning and co-workers, where *x* = D, T, Q, ..., yields the previously proposed "minimally augmented" basis sets, called maug-cc-pV(*x*+d)Z. Here, we present extensive and systematic tests of these basis sets for density functional calculations of chemical reaction barrier heights, hydrogen bond energies, electron affinities, ionization potentials, and atomization energies. The tests show that the maug-cc-pV(*x*+d)Z basis sets are as accurate as the aug-cc-pV(*x*+d)Z ones for density functional calculations, but the computational cost savings are a factor of about two to seven.

1. Introduction

For many quantum mechanical electronic structure calculations on molecules and chemical reactions, the results are sensitive to the inclusion of diffuse basis functions. Diffuse basis functions are spherical harmonics (or powers of Cartesian coordinates) times Gaussian functions with small exponents. These functions have long tails that allow the electrons to be farther from the nuclei. This is especially important for calculations on systems that require a good description of electrons in weakly bound orbitals or the outer parts of orbitals, such as many anions, transition states, and noncovalently bound systems.

Two systematic approaches to adding diffuse functions have emerged. The first is to add standard diffuse *s* and *p* functions (4 functions altogether) to nonhydrogenic atoms—this is called a "plus" or "+" basis set—or to add diffuse *s* and *p* basis functions to nonhydrogenic atoms and diffuse *s* basis functions to H and He—is called a "double +" or "++" basis set.[1] The second approach is to add a diffuse function to every atom for every symmetry already present in the original basis; this is called the augmented ("aug") approach.[2] For example, if a given basis set for sulfur atom has *s*, *p*, *d*, and *f* basis functions, one adds *s*, *p*, *d*, and *f* diffuse functions to that atom (a total of 16 functions, where all basis functions in this article use the spherical harmonic option—not the Cartesian one). Thus, as the underlying basis set becomes more complete, the number of diffuse functions increases. This makes the aug basis sets both larger and more rapidly convergent than the plus basis sets as the highest angular momentum of nondiffuse basis functions increases. However, in our previous paper,[3] we have shown that, in density functional calculations, the more expensive aug approach is not necessary; that is, the fixed number of diffuse functions of the plus sets is sufficient for results of double-, triple-, and quadruple-ζ quality. In particular, we showed that augmentation of the cc-pV*x*Z (where *x* = D, T, Q, ...) basis sets with the diffuse functions from the basis sets of Pople and co-workers, which yields cc-pV*x*Z+ basis sets, accounts for most of the effect that the full, much more expensive aug basis set provides. We also mentioned that the aug-cc-pV*x*Z basis sets can be truncated to contain only *s* and *p* diffuse functions on the non-hydrogenic atoms. We called this series of basis sets minimally augmented and abbreviated them as maug-cc-pV*x*Z, and we presented some calculations with this kind of basis set; however, the primary focus of the previous paper was on the plus strategy.

We have now carried out systematic tests of the performance of the maug-cc-pV*x*Z basis sets (where *x* stands for (D+d), (T+d), or (Q+d)) for density functional calculations, using both the popular B3LYP[4−7] density functional and also the recent, highly accurate M06-2X[8] density functional. We present the results of these tests here as a letter. Our tests involve computation of several commonly calculated and challenging molecular energetic properties. Barrier heights are the most important reaction parameters used for mechanism evaluation and kinetics calculations. Moreover, they are a good challenge for our purposes, since the accuracy of the description of transition states sometimes depends strongly on the presence and quality of the diffuse functions in a basis set. Since diffuse functions are often crucial in the description of noncovalent interactions such as hydrogen bonding, we also present tests on hydrogen bonding. Perhaps the most difficult test of the adequacy of the diffuse part of the basis set is the prediction of electron affinity values, because it involves anion calculations. Ionization potentials and atomization energy are also considered, because they are key thermochemical quantities.

In order to more specifically investigate the need for the diffuse basis functions on hydrogen atoms, we calculated electron affinity values for metal hydride.

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

10.1021/ct900566x
Published on Web 02/16/2010

Table 1. Definitions of the Diffuse Spaces in the Various Basis Sets, Numbers of Basis Functions in the Databases, Normalized and Relative Numbers of Basis Functions Raised to the Fourth Power, and the Normalized and Relative Computational Time of a Single-Point Energy Calculation on $C_4H_{10}S_2$ Using M06-2X

basis set	Li−Ar	H−He	N	$(N^4)_{Nor}$	$(N^4)_{Rel}$	$(C_4H_{10}S_2)_{Nor}$	$(C_4H_{10}S_2)_{Rel}$
aug-cc-pV(Q+d)Z	*spdfg*	*spdf*	17597	934	1.00	411	1.00
maug-cc-pV(Q+d)Z	*sp*		12665	251	0.27	74	0.18
cc-pV(Q+d)Z			12073	207	0.22	53	0.13
aug-cc-pV(T+d)Z	*spdf*	*spd*	10035	99	1.00	36	1.00
maug-cc-pV(T+d)Z	*sp*		7209	26	0.27	11.5	0.32
cc-pV(T+d)Z			6609	19	0.19	8.6	0.24
aug-cc-pV(D+d)Z	*spd*	*sp*	4989	6	1.00	4.5	1.00
maug-cc-pV(D+d)Z	*sp*		3783	2	0.33	1.5	0.32
cc-pV(D+d)Z			3183	1	0.17	1.0	0.22

To warn readers that all conclusions about the need for diffuse functions in density functional theory (DFT) calculations cannot be extended with equal success to wave function theory (WFT) calculations, we also present results of the electron affinity calculations at the second-order perturbation theory (MP2)[9] level.

Similar truncations as in maug-cc-pV(x+d)Z basis sets were performed for the maug-cc-pVxZ basis sets. The conclusions about the diffuse functions are the same as for the (x+d) series, which contains tight *d* functions on the elements in the 3p block of the periodic table. However, the use of (x+d) basis sets is recommended as they provide better quality results, especially at the DZ level and for hypervalent molecules. The results for aug-cc-pVxZ, maug-cc-pVxZ, and cc-pVxZ basis sets are available in the Supporting Information, and the main body of this letter will discuss only aug-cc-pV(x+d)Z, maug-cc-pV(x+d)Z, and cc-pV(x+d)Z basis sets.

2. Methods and Databases

The databases employed for the present study are for barrier heights, hydrogen-bond interaction energies, electron affinities, ionization potentials, and bond energies. All energetic results presented in this communication were obtained using the *Gaussian 03*[10] program and the MN-GFM[11] functional module. The cost estimates listed in Table 1 were calculated using the *Gaussian 09* program. All results are based on single-point calculations run at geometries[12] optimized with the QCISD/MG3 method (for the BH24, EA13/3, IP13/3, and AE6 databases) and the MC-QCISD/3[13] method (for the HB6 database), and vibrational contributions are excluded (that is, we are testing the methods for Born−Oppenheimer electronic energies (including nuclear repulsion), not for enthalpies). QCISD denotes quadratic configuration interaction with single and double excitations.[14] The geometries for all the species in the five databases considered are available in the databases' respective references given below.

The results provided by the two density functionals B3LYP and M06-2X with the fully augmented aug-cc-pV(x+d)Z basis sets, minimally augmented maug-cc-pV(x+d)Z basis sets, and unaugmented cc-pV(x+d)Z basis sets containing no diffuse functions were compared to the best estimates in the following databases:

- DBH24/08[15,16] database of diverse barrier heights, consisting of the best estimates for 24 barrier heights for the heavy-atom transfer reaction, nucleophilic substitution, hydrogen transfer, and unimolecular and association reactions
- HB6[17] database, consisting of hydrogen bond energies for $(NH_3)_2$, $(HF)_2$, $(H_2O)_2$, $NH_3(H_2O)$, $(HCONH_2)_2$, and $(HCOOH)_2$
- EA13/3[18,19] database, consisting of electron affinities for C, S, O, Si, P, Cl, OH, PH, SH, PH_2, O_2, S_2, and Cl_2
- IP13/3[18,19] database, consisting of ionization potentials for the same 13 species as in the EA13/3 database
- AE6[20] representative atomization energy database, consisting of atomization energies of SiH_4, SiO, S_2, propyne, glyoxal, and cyclobutane.

Since the computed values are compared here with experimental or well converged theoretical values in the databases, all the computed energetic data had the spin−orbit contributions added on for F, C, O, Cl, Si, S, OH, and HS,[21] and the experimental data had the zero-point contributions subtracted from them.

In Table 1, we define the diffuse space in the fully augmented, minimally augmented, and unaugmented basis sets. In order to compare the approximate cost of the calculations involving the basis sets used in the tests, we report the sum N of the number of basis functions used in the calculations, summed over all the test cases in all five databases used, and the sum raised to the fourth power (N^4), which is how the cost of the hybrid DFT calculations in popular computer programs scale in the limit of large systems, when linear-scaling algorithms are not used. For clarity, the N^4 values were normalized (denoted by subscript Nor) to the N^4 value for the least expensive basis set cc-pV(D+d)Z. In order to quantify the cost savings for a given x (x = D, T, Q) achieved by using maug basis sets instead of aug ones, we show in the $(N^4)_{Rel}$ column values normalized to the aug-cc-pV(x+d)Z N^4 values for each x. Since the limit of asymptotic scaling is never fully reached in practice, we also illustrate the timings with a real example; in particular, we list timings of single-point energy calculations on the medium-size molecule 1,4-butanedithiol ($C_4H_{10}S_2$).

3. Results

Tables 2−7 provide the mean signed errors (MSEs) and the mean unsigned errors (MUEs, which can also be called mean absolute errors) for a given level of theory in the calculations involving species in a given database. We define MSE and MUE as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{1}$$

$$\mathrm{MUE} = \frac{1}{n}\sum_{i=1}^{n} |e_i| \tag{2}$$

In eqs 1 and 2, e_i is an error in a single property value (for example, an ionization potential value for one molecule) in the database, which contains a total of n such values for different

Table 2. Errors in Predictions of the Barrier Heights (kcal/mol) in the BH24/08 Database

	B3LYP		M06-2X	
basis set	MSE	MUE	MSE	MUE
aug-cc-pV(Q+d)Z	−3.99	4.06	0.03	0.93
maug-cc-pV(Q+d)Z	−3.97	4.06	0.09	0.92
cc-pV(Q+d)Z	−4.79	4.85	−0.50	1.38
aug-cc-pV(T+d)Z	−4.08	4.14	−0.06	0.88
maug-cc-pV(T+d)Z	−4.02	4.10	0.08	0.91
cc-pV(T+d)Z	−5.26	5.42	−0.87	2.06
aug-cc-pV(D+d)Z	−4.83	4.90	−0.60	1.20
maug-cc-pV(D+d)Z	−4.69	4.72	−0.39	1.33
cc-pV(D+d)Z	−7.01	7.54	−2.32	3.75

Table 3. Errors in Predictions of the Hydrogen Bonding Energies (kcal/mol) in the HB6 Database

	B3LYP		M06-2X	
basis set	MSE	MUE	MSE	MUE
aug-cc-pV(Q+d)Z	−0.76	0.76	0.11	0.29
maug-cc-pV(Q+d)Z	−0.73	0.73	0.10	0.28
cc-pV(Q+d)Z	−0.23	0.55	0.42	0.46
aug-cc-pV(T+d)Z	−0.73	0.73	0.17	0.31
maug-cc-pV(T+d)Z	−0.67	0.67	0.17	0.34
cc-pV(T+d)Z	0.51	0.74	1.02	1.02
aug-cc-pV(D+d)Z	−0.39	0.40	0.37	0.37
maug-cc-pV(D+d)Z	−0.27	0.64	0.43	0.61
cc-pV(D+d)Z	2.97	2.97	2.91	2.91

Table 4. Errors in Predictions of the Electron Affinities (kcal/mol) in the EA13/3 Database

EA13/3	B3LYP		M06−2X	
basis set	MSE	MUE	MSE	MUE
aug-cc-pV(Q+d)Z	−1.99	2.33	0.78	1.47
maug-cc-pV(Q+d)Z	−1.96	2.31	0.82	1.49
cc-pV(Q+d)Z	4.35	5.30	5.30	5.30
aug-cc-pV(T+d)Z	−2.07	2.37	0.67	1.46
maug-cc-pV(T+d)Z	−2.04	2.36	0.70	1.49
cc-pV(T+d)Z	9.49	10.13	9.99	9.99
aug-cc-pV(D+d)Z	−2.45	2.75	0.87	2.21
maug-cc-pV(D+d)Z	−2.53	2.92	0.71	2.28
cc-pV(D+d)Z	20.59	20.85	20.37	20.37

Table 5. Errors in Predictions of the Ionization Potentials (kcal/mol) in the IP13/3 Database

	B3LYP		M06-2X	
basis set	MSE	MUE	MSE	MUE
aug-cc-pV(Q+d)Z	3.42	4.62	0.66	2.26
maug-cc-pV(Q+d)Z	3.41	4.61	0.64	2.25
cc-pV(Q+d)Z	3.19	4.42	0.51	2.14
aug-cc-pV(T+d)Z	3.54	4.65	0.99	2.63
maug-cc-pV(T+d)Z	3.53	4.63	0.96	2.62
cc-pV(T+d)Z	2.94	4.14	0.57	2.31
aug-cc-pV(D+d)Z	4.01	4.72	1.08	2.81
maug-cc-pV(D+d)Z	4.14	4.67	1.26	2.99
cc-pV(D+d)Z	1.67	3.15	−0.67	2.84

species. The mean error values for the atomization energies of the molecules in AE6 are divided by 4.83, which is the average number of bonds per molecule in this database, so that the results may be interpreted on a per bond basis.

It may be useful to comment on the meaning of the electron affinities. It is well known that density functional calculations with certain kinds of approximate density functionals predict

Table 6. Errors in Predictions of the Atomization Energies (kcal/mol per bond) in the AE6 Database

	B3LYP		M06-2X	
basis set	MSE	MUE	MSE	MUE
aug-cc-pV(Q+d)Z	−0.42	0.59	0.01	0.22
maug-cc-pV(Q+d)Z	−0.43	0.61	0.01	0.22
cc-pV(Q+d)Z	−0.38	0.56	0.03	0.22
aug-cc-pV(T+d)Z	−0.68	0.76	−0.19	0.33
maug-cc-pV(T+d)Z	−0.70	0.79	−0.23	0.36
cc-pV(T+d)Z	−0.59	0.69	−0.16	0.29
aug-cc-pV(D+d)Z	−2.31	2.31	−1.65	1.73
maug-cc-pV(D+d)Z	−2.67	2.67	−1.94	1.94
cc-pV(D+d)Z	−2.29	2.29	−1.59	1.65

Table 7. Electron Affinity (kcal/mol) of Lithium Hydride LiH

basis set	B3LYP	M06-2X
aug-cc-pV(Q+d)Z	10.12	6.55
jul-cc-pV(Q+d)Z	10.14	6.52
maug-cc-pV(Q+d)Z	10.05	6.42
cc-pV(Q+d)Z	7.82	3.71
aug-cc-pV(T+d)Z	9.97	6.28
jul-cc-pV(T+d)Z	9.96	6.20
maug-cc-pV(T+d)Z	9.84	6.09
cc-pV(T+d)Z	6.24	1.96
aug-cc-pV(D+d)Z	10.33	6.51
jul-cc-pV(D+d)Z	10.20	6.18
maug-cc-pV(D+d)Z	10.05	6.18
cc-pV(D+d)Z	5.75	1.15

unbound negative results in the limit of a large basis set.[22] Nevertheless, calculations with standard basis sets have been shown to often give stable and useful results.[23,24] The results presented here are a test of the stability of such standard calculations to the size of the diffuse subspace of the basis set.

4. Discussion

The tables show that there is essentially complete agreement between predictions of the fully (aug) and minimally (maug) augmented basis sets for density functional calculations in the triple-ζ and quadruple-ζ cases. For quadruple-ζ, the MUE of the maug results is actually lower than the MUE of the aug results in six of the ten cases, and in the other four cases, the maug MUE is never higher than the aug MUE by more than 3%. And yet, the diffuse functions have a very important effect. For example, in Table 4 for electron affinities, the MUE for the unaugmented quadruple-ζ basis is a factor of 2 or 4 times higher than the error in the maug basis, but no significant increase in accuracy is attained by proceeding from the maug basis to the fully augmented basis. At the triple-ζ level, the effects of diffuse functions are larger, for example, decreasing the MUE by factors of 4 and 7 for the electron affinities, but again the maug MUE is almost the same as the aug MUE. For triple-ζ, maug has a lower MUE for five of the ten cases and never has a MUE higher than the aug one by more than 9%.

The effect of diffuse functions is largest for double-ζ basis sets, with the MUE for electron affinities in unaugmented calculations being 7 or 9 times larger than that for the maug basis. We believe it is due to the fact that the unaugmented cc-pVDZ basis set is the least diffuse basis set considered in the present article. Therefore, the effect of pruning of some of the diffuse functions on the quality of the results is the most significant at the double-ζ level. On average though, the ratio

Table 8. Errors in Predictions of the Electron Affinity (kcal/mol) in the EA13/3 Database in MP2 Calculations

basis set	MSE	MUE
aug-cc-pV(T+d)Z	1.44	2.26
maug-cc-pV(T+d)Z	4.01	4.31
cc-pV(T+d)Z	14.36	14.36

of the maug MUE to the aug one is greater than unity by only 13%, and it is only 2% greater than unity if we omit hydrogen bonding. The full aug set of the diffuse functions decreases the double-ζ MUE for hydrogen bonding by a factor of 7−8, whereas the maug diffuse functions decrease it by only a factor of 5.

The cost savings that come from using the maug series of basis sets instead of aug are very large. Using the N^4 asymptotic scaling factors of Table 1 shows that the minimally augmented diffuse basis sets maug-cc-pV(x+d)Z offer the same quality results as their aug analogs at a cost reduced by 73%, 73%, and 67% for x = Q, T, and D, respectively. Alternatively if we use actual costs for specific calculations on 1,4-butanedithiol, the time savings using maug vs aug basis sets are 82%, 68%, and 68% for quadruple, triple, and double-ζ basis sets, respectively, and for the B3LYP functional (these timings not shown in the table), they are 86%, 76%, and 52%.

One caveat on the conclusions drawn here is that the databases used for the tests presented here include no metal atoms, so the conclusions have been established only for nonmetals, although it would not be surprising if they were also found to hold for compounds containing metal atoms.

Another noteworthy point is that maug triple-ζ is usually very close to the basis set limit for DFT, at least for the nonmetal systems in the present study.

To show that the diffuse functions on hydrogen are practically redundant, we present the dependence on the electron affinity of lithium hydride (LiH) in Table 7. A metal hydride would be the case where H would be most likely to need diffuse functions, and electron affinities provide the toughest test of the need for diffuse functions, so the particular choice of an electron affinity of a metal hydride is a serious challenge. Deleting the diffuse functions on hydrogenic atoms from an "aug-" basis set yields what we call a "jul-" basis set. From the data shown in Table 7, one can see that even for the electron affinity of metal hydride the diffuse functions on hydrogen are unnecessary for DFT calculations of energetic molecular properties. In particular, the difference in performance of aug and jul basis sets is practically nonexistent. Results for two other metal hydrides (BeH and MgH) are available in the Supporting Information; the error encountered by using jul instead of aug varies from 0.2% to 3.2% and from 1.4% to 9.4% for MgH and BeH, respectively, and so the results for these systems confirm the unimportance of diffuse functions on hydrides. For completeness, we note that the monatomic hydrogen anion is an exception; for this one-center, two-electron system, it is essential to include diffuse functions for an accurate description (however, since the energy of this system has already been calculated accurately to several significant figures with explicitly correlated basis functions, its basis set requirements are a not a major concern here).

Table 8 shows that WFT is more slowly convergent than DFT with respect to the number of the diffuse functions on heavy atoms in the basis set used. However, the effect of diffuse functions on H has previously been shown to be negligible for energetic calculations even in WFT. In particular, an extensive study of basis set effects on the calculated bond energy and electron affinity of LiH showed that "diffuse functions on hydrogen have little importance for thermochemical calculations."[25]

The present study adds to previous work showing that conclusions about basis sets derived from many years of experience with WFT calculations do not necessarily hold for DFT.[18,26−31]

Although it is not the main point of this paper, we note that Tables 2−6 contain 45 direct comparisons of mean unsigned errors for B3LYP to those for M06-2X for a given basis set and database. In one case, the mean unsigned errors are the same, and in all other cases, M06-2X has the better performance.

5. Conclusions

The only case in Tables 2−6 where the augmentation of the cc-pV(x+d)Z basis sets with the full aug set of the diffuse functions performs significantly better in the density functional theory calculations than the maug basis set is for two studied cases of hydrogen bonding at the double-ζ level. In the other 28 cases considered here, the maug basis actually has a lower mean unsigned error (MUE) than the aug one in 12 of the cases, and the average ratio of the maug MUE to the aug one is only 1% greater than unity.

The present tests have been restricted to the main group. We recommend that, for energetic molecular properties, including barrier heights, the aug basis sets be truncated to the maug level for density functional calculations on systems composed of main-group atoms.

Acknowledgment. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, under grant no. DE-FG02-86ER13579.

Supporting Information Available: Additional tables showing root-mean-square errors and results for basis sets without tight d functions plus calculations on MgH and BeH. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.

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

(3) Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1197. Erratum and addendum: Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 3330.

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

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

(6) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5848.

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

(8) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215. Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(9) Mólller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(10) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(11) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module*, version 3.0; University of Minnesota: Minneapolis, MN, 2007.

(12) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.

(13) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898.

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

(15) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 569.

(16) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808.

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

(18) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.

(19) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898.

(20) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *107*, 8996. Erratum: *J. Phys. Chem. A* **2004**, *108*, 1460.

(21) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.

(22) Jarecki, A. A.; Davidson, E. R. *Chem. Phys. Lett.* **1999**, *300*, 44.

(23) Galbraith, J. M.; Schaefer, H. F., III. *J. Chem. Phys.* **1996**, *105*, 862.

(24) Horny, L.; Petraco, N. D. K.; Schaefer, H. F., III. *J. Am. Chem. Soc.* **2002**, *124*, 14716.

(25) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *ACS Symp. Ser.* **2007**, *958*, 153.

(26) Jensen, F. *J. Chem. Phys.* **2002**, *116*, 7372. Jensen, F. *J. Chem. Phys.* **2002**, *117*, 9234.

(27) Prascher, B. P.; Wilson, A. K. *Mol. Phys.* **2007**, *105*, 2899.

(28) Schneider, A. C.; Andzelm, J. W. *J. Comput. Chem.* **1997**, *18*, 775.

(29) Halls, M. D.; Schlegela, H. B., *J. Chem. Phys.* **1998**, *109*, 10587.

(30) Florian, J.; Johnson, B. G. *J. Phys. Chem.* **1995**, *99*, 5899.

(31) Bauschlicher, C. W.; Partridge, H., Jr. *Chem. Phys. Lett.* **1995**, *240*, 533.

CT900566X

JCTC Journal of Chemical Theory and Computation

Single Electron Transfer and S_N2 Reactions: The Importance of Ionization Potential of Nucleophiles

Ronald R. Sauers*

Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08903

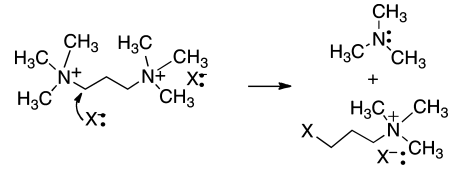
Received November 19, 2009

Abstract: The importance of single electron transfer energetics in promoting S_N2 reactions was probed by a density functional computational study on substitution reactions of quaternary ammonium ion complexes with anionic nucleophiles. Good correlations were found between the ionization potentials (IP) of the nucleophiles when plotted against density functional theory (DFT)-computed reaction activation enthalpies ($\Delta H^{\neq}_{rxn}$) over a range of 15 eV. Poor correlations were found between IPs or proton affinities and central barrier heights ($\Delta H^{\neq}_{cmp}$). Examples of inverted values of $\Delta H^{\neq}_{rxn}$ of primary vs secondary systems were found.

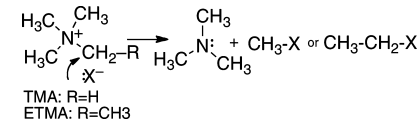
Introduction

Bimolecular nucleophilic substitution reactions at saturated carbon atoms likely represent the most highly studied organic mechanism both experimentally and computationally. The pioneering work of the Ingold school established the bimolecular nature of the reaction in solution.[1] The work of Brauman et al.[2] laid the theoretical groundwork for study of gas phase S_N2 reactions in terms of a double minimum potential energy profile to form ion−molecule complexes that either revert to starting components or pass over "central" S_N2 transition energy barriers. The reaction products also form a complex that ultimately dissociates to individual products. Recent work by McMahon and co-workers has provided direct experimental verification of these ideas.[3] Bento and Bickelhaupt have carried a detailed analysis of methyl halide/halide reactions, Si, and group 14 elements and have concluded that HOMO/LUMO orbital interactions are of paramount importance.[4] Shaik and others[5] have stressed the importance of single electron transfer from the nucleophile to the substrate as an important feature of the

Scheme 1. Nucleophilic Substitution by Halogen on a bis-Quaternary Ammonium Salt



Scheme 2. Reactions of TMA/ETMA with Nucleophile X^-



electronic nature of S_N2 transition states. In this context, the transition state for simple S_N2 reactions involving a single electron transfer can be symbolized as shown below.

$$[^{-}Nu^{\bullet}R^{\bullet\bullet}X \leftrightarrow Nu^{\bullet\bullet}R^{\bullet}X^{-}]^{\neq}$$

This study probes the idea that energies of S_N2 transition structures should correlate with ionization potentials (IPs) of attacking nucleophiles.

This work evolved from experimental and computational studies with bis-ammonium alkyl dihalides that were shown to undergo substitution and elimination reactions in the gas phase (Scheme 1).[6] An unexpected result was the relative insensitivity of the computed central barrier ($\Delta H^{\neq}_{cmp}$) to the nature of the halogen. It was also shown that the second positive center had little effect on the computed S_N2 activation enthalpies compared to simple tetraalkylammonium ion salts. These results stimulated a desire to extend the survey of S_N2 reactivity to a much broader range of nucleophiles and in particular to assess the relationship between nucleophile ionization potential on transition structure energy and S_N2 barriers. Scheme 2 describes this approach that compares the energetics of reactions of nucleophilies at primary and secondary carbon centers of ethyltrimethyl (ETMA) and tetramethyl (TMA) ammonium salts with trimethylamine as the common leaving group.

The abbreviated enthalpy diagram in Scheme 3 outlines the computational analysis invoked in this study given the expectation that complex formation between neutral products trimethylamine and alkylX would not affect the overall objectives. ΔH_{pr} represents the sum of the enthalpies of R−X and trimethyl amine.

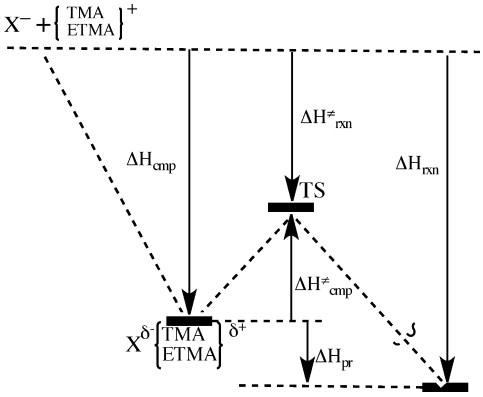
Computations and Methodology

All structures were fully optimized by analytical gradient methods using the Gaussian03 suite[7] and DFT calculations at

* E-mail: sauers@rutchem.rutgers.edu.

10.1021/ct900611g
Published on Web 02/04/2010

Scheme 3. Enthalpy Diagram for Nucleophilic Substitution Reactions of the Tetramethylammonium Ion (TMA) and Ethyltrimethylammonium Ion (ETMA) with Nucleophile X^-



the RB3LYP/6-31+G(d) level,[8] the exchange functional of Becke,[9] and the correlation functional of Lee, Yang, and Parr.[10] Vibrational analyses established the nature of all stationary points as either energy minima (no imaginary frequencies) or first-order transition structures (one imaginary frequency). In some cases, IRC calculations verified that the TS connected intitial and product structures. In other cases, we relied on animations of the vectors associated with the imaginary frequencies as a guide. Reported enthalpies (unscaled) were corrected for zero-point energy and temperature effects at 298.14 K. In the first stage of this project, the enthalpies of complex formation (ΔH_{cmp}) between the anions and TMA and ETMA were evaluated. S_N2 transition structures ($\Delta H^{\neq}_{cmp}$) were located for reactions at methyl groups in the TMA series and at the methylene groups in the ETMA series, i.e., reactions at primary and secondary carbons with trimethylamine as a leaving group (Scheme 2). These data are summarized in Table 1 along with derived values for the transition structure enthalpies $\Delta H^{\neq}_{rxn}$.

Because few experimental IPs for many of the ions are available, they were computed using DFT at the (U)B3LYP/aug-cc-pVDZ level. Computed values agreed within 2−3 kcal/mol with available experimental values. Proton affinities were compiled from the literature or computed. Table 2 displays the relevant ionization energies and the proton affinities.

Table 2. Ionization Potentials[a] and Proton Affinities:[b] (U)B3LYP/aug-cc-pVDZ[c]

anion	ionization potential, eV	proton affinity, kcal mol^{-1}	anion	ionization potential, eV	proton affinity, kcal mol^{-1}
F^-	3.56	371.3	OCl^-	2.34	349.4
Cl^-	3.72	333.4	CO_3^{-2}	−3.55	476.5
HCO_2^-	3.48	345.3	H^-	0.89	394.6
NO_3^-	3.90	327.5	$CH_3CH_2O^-$	1.67	378.6
OH^-	1.85	390.8	$HOBO_2^{-2}$	−2.52	486.9
CH_3O^-	1.53	381.7	BO_3^{-3}	−8.38	592.8
CN^-	4.04	351.1	$(HO)_2PO_2^-$	4.15	325.4
HOO^-	1.03	368.5	$HOPO_3^{-2}$	−1.84	452.7
N_3^-	2.67	337.6	PO_4^{-3}	−7.21	573.1
SH^-	2.36	351.1	$HSiO_4^{-3}$	−7.13	573.5
$HOCO_2^-$	3.56	330.4	SiO_4^{-4}	−11.1	664.7

[a] DFT methodology has been shown to give good results for IPs: Rienstra-Kiracofe, J. C.; Tschumper, G. S.; Schaefer, H. F., III. Atomic and Molecular Electron Affinities: Photoelectron Experiments and Theoretical Computations. *Chem. Rev.* **2002**, *102*, 231−282. [b] Some proton affinity data is taken from tables in the NIST data bank: Bartmess, J. E. In *NIST Standard Reference Database Number 69*; Mallard, W. G., Linstrom, P. J., Eds.; National Institutes of Standards and Technology (http://webbook.nist.gov): Gaithersburg, MD, 1999. [c] See Supporting Information Table S5 for primary data.

Discussion

Figure 1 shows a plot of central barrier heights ($\Delta H^{\neq}_{cmp}$) for the S_N2 reactions of ETMA vs IPs. The widely scattered data points reveal no meaningful correlations.[11] Similar results were obtained for reactions of TMA (not shown). In part, these findings reflect the consequences of the small range of $\Delta H^{\neq}_{cmp}$ compared to the large variations in IP. Likewise, plots of $\Delta H^{\neq}_{cmp}$ vs proton affinities (PA) for ETMA (Figure 2; and TMA)

Table 1. Computed Data for Complex Energy Formation (ΔH_{cmp}), Transition Structure Barriers ($\Delta H^{\neq}_{cmp}$), and Reaction Transition Structure Enthalpy Changes ($\Delta H^{\neq}_{rxn}$) for ETMA and TMA Systems at 298.15 K: B3LYP/6-31+G(d)

	tetramethylammonium systems			ethyltrimethylammonium systems		
anion	ΔH_{cmp}, kcal mol^{-1}	$\Delta H^{\neq}_{cmp}$, kcal mol^{-1}	$\Delta H^{\neq}_{rxn}$, kcal mol^{-1}	ΔH_{cmp}, kcal mol^{-1}	$\Delta H^{\neq}_{cmp}$, kcal mol^{-1}	$\Delta H^{\neq}_{rxn}$, kcal mol^{-1}
F^-	−114.33	22.3	−92.02	−112.75	22.6	−90.15
Cl^-	−91.98	19.4	−75.59	−90.66	21.5	−69.12
HCO_2^-	−95.28	27.4	−67.89	−93.44	28.1	−65.34
NO_3^-	−85.15	21.2	−63.92	−83.70	21.2	−62.50
OH^-	−117.25	24.1	−93.20	−111.40	19.3	−92.10
CH_3O^-	−107.94	19.5	−88.43	−102.40	17.1	−85.30
CN^-	−88.69	18.1	−70.60	−83.77	19.3	−64.47
HOO^-	−109.10	16.7	−92.39	−107.65	17.2	−90.45
N_3^-	−85.87	16.2	−69.68	−84.56	16.7	−67.86
SH^-	−89.65	14.8	−74.85	−85.24	14.3	−70.94
$HOCO_2^-$	−92.46	19.8	−72.68	−88.83	18.1	−70.73
OCl^-	−99.58	16.4	−83.16	−95.24	14.3	−80.94
CO_3^{-2}	−192.33	19.2	−173.12	−186.72	12.2	−174.53
H^-	−142.10	14.0	−128.06	−140.36	15.4	−124.97
$CH_3CH_2O^-$	−104.92	12.6	−92.35	−99.37	16.9	−82.47
$HOBO_2^{-2}$	−200.64	21.8	−178.82	−203.85	22.9	−180.96
BO_3^{-3}	−314.87	20.0	−294.86	−311.44	14.2	−297.27
$(HO)_2PO_2^-$	−88.88	23.6	−65.23	−85.87	21.8	−64.05
$HOPO_3^{-2}$	−178.86	19.3	−159.52	−174.09	13.6	−160.48
PO_4^{-3}	−282.44	22.1	−260.38	−280.22	11.3	−268.95
$HSiO_4^{-3}$	−289.90	23.5	−266.44	−280.66	11.7	−268.94
SiO_4^{-4}	−419.02	19.4	−399.57	−418.38	15.3	−403.08

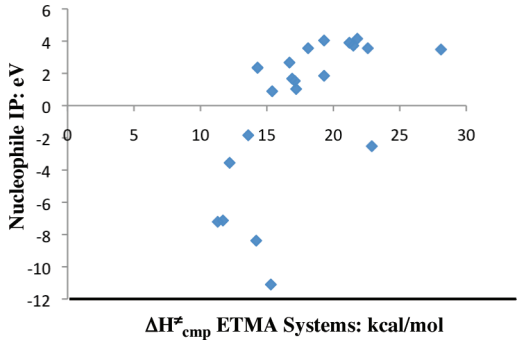


Figure 1. Plot of IP vs $\Delta H^{\neq}_{cmp}$ for ETMA.

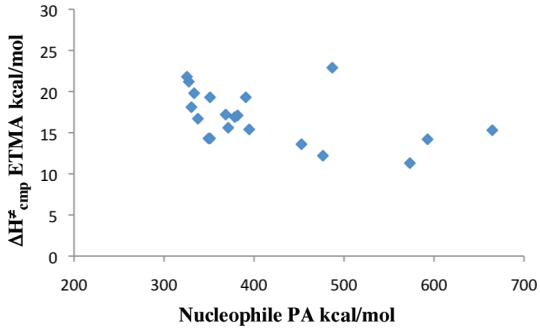


Figure 2. Plot of $\Delta H^{\neq}_{cmp}$ vs PA for ETMA.

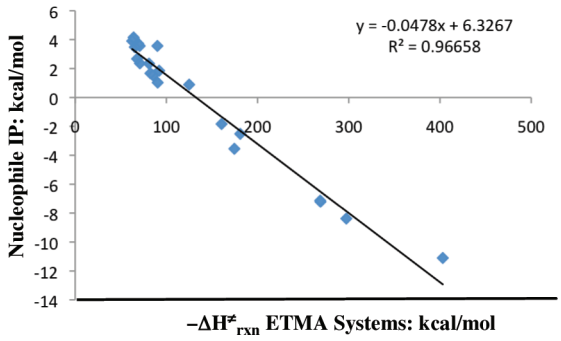


Figure 3. Plot of IP of nucleophiles vs $-\Delta H^{\neq}_{rxn}$ for tetramethylammonium systems.

systems gave rise to widely scattered arrays of data points. Similarly, attempts to find significant correlations for plots of $-\Delta H^{\neq}_{cmp}$ of monoanions vs IP and plots of $-\Delta H^{\neq}_{cmp}$ vs IP for oxygen-centered nucleophiles failed. On the other hand, plots of $-\Delta H^{\neq}_{rxn}$ vs IP gave good correlations for both methyl and ethyl substitutions (Figures 3 and 4).

These remarkable trends are consistent with the concept of single electron transfer as an important electronic factor in initiating bimolecular nucleophilic substitution reactions at primary and secondary saturated carbon atoms. In addition, these correlations validate the concept that it is the stability of the transition structure relative to the separate reactants that is the kinetically relevant term as opposed to central barrier heights, per se.[12]

Some unexpected results were the findings that $\Delta H^{\neq}_{cmp}$ for substitution at the methyl carbon was not always more favorable than that for substitution at the ethyl carbon (Table 1, Figure 5). These unprecedented discrepancies showed up not only with the polyionic anions but also with simple nucleophiles, e.g., hypochlorite ion, hydroxide ion, et al. That these disparities are

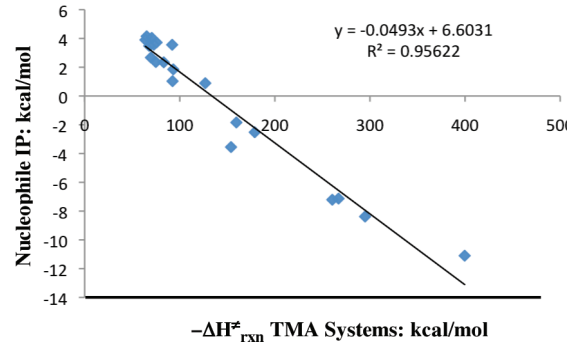


Figure 4. Plot of IP of nucleophiles vs $-\Delta H^{\neq}_{rxn}$ for ethyltrimethylammonium systems.

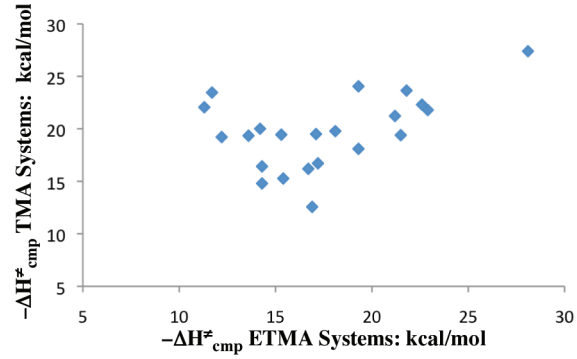


Figure 5. Plot of $-\Delta H^{\neq}_{cmp}$: TMA vs ETMA.

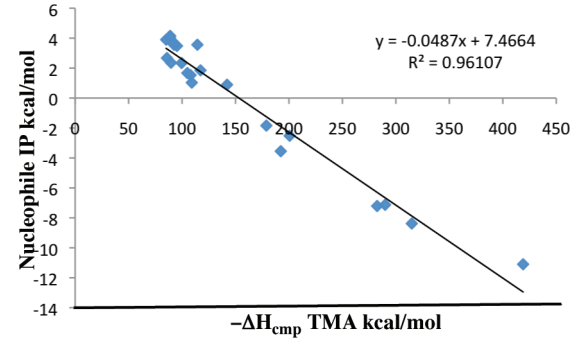


Figure 6. Plot of IP of nucleophiles vs $-\Delta H_{cmp}$ for tetramethylammonium ions.

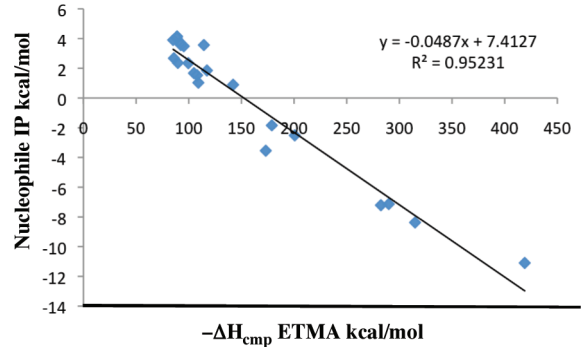


Figure 7. Plot IP of nucleophiles vs $-\Delta H_{cmp}$ for ethyltrimethylammonium ions.

not anomalies associated with complex formation is shown by plots of IP vs ΔH_{cmp} (Figures 6 and 7).

Apparently the electronic forces and steric and/or entropic[13] factors involved in complex formation vs the transition structures

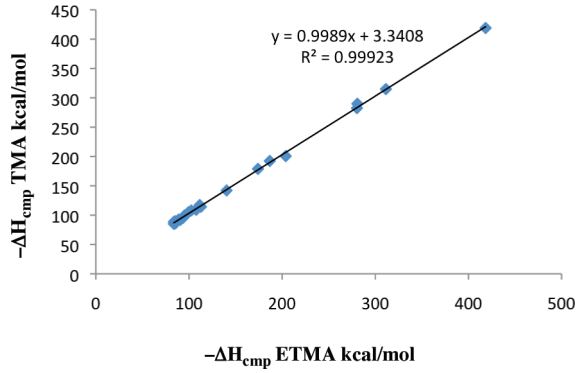


Figure 8. Comparison of $-\Delta H_{cmp}$: TMA vs ETMA.

for certain of these systems can be significantly disparate. Similarly, comparisons of the enthalpies of the complex formation did not show any major anomalies as shown by the data in Figure 8. An excellent correlation was obtained: in almost all cases, methyl systems give rise to tighter, lower enthalpy complexes than ethyl analogs. The only exception was HBO_3^{-2}: the TMA complex was less stable than the ETMA complex by ~3 kcal mol^{-1}. It would appear that irregularities in binding energies/structures do not give rise to anomalous methyl/ethyl reversals.[12,13] Comparisons of the TMA vs ETMA transition structures in general showed that the TMA transition structures were somewhat "earlier" in that average X···C and C···N bond lengths were shorter (1.774 and 2.254 Å, respectively) than those in the corresponding ETMA TSs (1.875 and 2.305 Å, respectively). No trends were noted that resolved the reactivity reversals, however (see Supporting Information Table S6).

Detailed studies by Bento and Bickelhaupt[4] with combinations of S_N2 reactions of methyl halide/halide ions support our results in that these workers describe nucleophilic character in terms of charge transfer and HOMO/LUMO interactions.

In related studies, Uggerud[14] recently estimated potential energy profiles for 18 identity S_N2 reactions using G2 quantum methods and found a linear correlation between nucleophile ionization potential (1−4 eV range) and barrier heights. It is interesting to note that the slope of the plot of IP vs $\Delta H^{\neq}_{cmp}$ was much steeper than the ones we found in Figures 3 and 4. This is probably due to the fact that identity reactions do not involve an overall energy change, whereas the systems studied herein are all strongly exothermic (Supporting Information Tables S4 and S5).

The alkyl ammonium systems examined in this study represent one extreme of structure−reactivity behavior in that the reactant bears a positive charge that is partially neutralized in the transition structure unlike most of the previous studies. For example, Brauman and Olmstead found that the cyanide ion is virtually unreactive toward methyl chloride.[15] In this study, $\Delta H^{\neq}_{rxn}$ for CN^- is not atypical compared to other unicharged nucleophiles, and the transition structure has a "typical" geometry.

In many cases, the high degree of exothermicity of these reactions, −61 to −400 kcal mol^{-1} (see Supporting Information Tables S4 and S5), may serve to compress the range of activation enthalpies because the geometry of the transition structures resides close to that of the starting components (vide infra) in contrast to "identity S_N2 reactions" in which the transition structure involves equal bonding of the nucleophile and leaving group. Still, there are some exceptions and inconsistencies with this argument. The TSs for attack of the nitrate ion on both methyl and ethyl centers are nearly symmetrical, and the $\Delta H^{\neq}_{cmp}$'s are essentially the same. Yet, in both cases, the overall reactions are highly exothermic: 88.8 vs 91.1 kcal mol^{-1}, respectively.

In a recent computational study of reactions of substituted p-X-phenoxides with methyl halides, Li and Xue[16] found good correlations of both barrier heights and central barriers with the phenoxide substituent (σ) constants. These workers did not examine correlations with ionization potential, but it is likely that IPs would parallel nucleophilicity of the phenoxides.

Conclusions

The reported correlations of IP with transition structure energy serve as a useful predictor of $\Delta H^{\neq}_{rxn}$.[17] Clearly, there are subtleties, e.g., entropy, steric effects, and orbital shapes, that modulate the behavior of the complexes and that are not understood at this time. It is clear that barrier heights $\Delta H^{\neq}_{cmp}$ are not useful predictors of relative reactivity in these systems. The generality of these correlations and exceptions remains to be explored in other systems.[4]

Acknowledgment. We thank Profs. K. Krogh-Jespersen, J. K. Lee, and H. Haubenstock for helpful comments and suggestions.

Supporting Information Available: Transition structure bond distances, ΔH_{rxn} data, ΔH_{pr} data, ionization potential calculations, and enthalpy data for all reactions. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) The kinetics of the S_N2 reaction were first developed with substitution reactions on quaternary ammonium salts: Hughes, E. D.; Ingold, C. K.; Patel, C. S. Influence of Poles and Polar Linkings on the Course Pursued by Elimination Reactions. Part XVI. Mechanism of the Thermal Decomposition of Quaternary Ammonium Compounds. *J. Chem. Soc.* **1933**, 526–530.

(2) Brauman, J. I.; Olmstead, W. N.; Lieder, C. A. Gas-phase nucleophilic displacement reactions. *J. Am. Chem. Soc.* **1974**, *96*, 4030–4031.

(3) Li, C.; Ross, P.; Szulejko, J. E.; McMahon, T. B. High-Pressure Mass Spectrometric Investigations of the Potential Energy Surfaces of Gas-Phase S_N2 Reactions. *J. Am. Chem. Soc.* **1996**, *118*, 9360–9367.

(4) Bento, A. P.; Bickelhaupt, F. M. Nucleophilicity and Leaving-Group Ability in Frontside and Backside S_N2 Reactions. *J. Org. Chem.* **2008**, *73*, 7290–7299. Bento, A. P.; Bickelhaupt, F. M. Nucleophilic Substitution at Silicon (S_N2@Si) via a Central Reaction Barrier. *J. Org. Chem.* **2007**, *72*, 2201–2207. Bento, A. P.; Bickelhaupt, F. M. Frontside versus Backside S_N2 Substitution at Group 14 Atoms: Origin of Reaction Barriers and Reasons for Their Absence. *Chem. Asian J.* **2008**, *3*, 1783–1792.

(5) (a) Shaik, S. S. The Collage of S_N2 Reactivity Patterns: A State Correlation Diagram Model. *Prog. Phys. Org. Chem.* **1985**, *15*, 197–337. (b) Pross, A.; Shaik, S. S. A qualitative valence-bond approach to organic reactivity. *Acc. Chem. Res.* **1981**, *16*, 363–370. (c) Shaik, S. S.; Schlegel, H. B.; Wolfe, S. *Theoretical Aspects of Physical Organic Chemistry: The S_N2 Mechanism*; Wiley: New York, 1992. (d) Pross, A. A qualitative valence-bond approach to organic reactivity. *Acc. Chem. Res.* **1985**, *18*,

212–219. (e) Shaik, S. S. The S_N2 and Single Electron Transfer Concepts. A Theoretical and Experimental Overview. *Acta Chem. Scand.* **1990**, *44*, 205–21. (f) Saveant, J. M. *Advances in Physical Organic Chemistry* **1990**, *26*, 1–130.

(6) Aime, C.; Plet, B.; Manet, S.; Schmitter, T.-M.; Hue, I.; Oda, R.; Sauers, R. R.; Romsted, L. S. Competing Gas Phase Substitution and Elimination Reactions of Gemini Surfactants with Anionic Counterions by Mass Spectrometry. Density Functional Theory Correlations with Their Bolaform Halide Salt Models. *J. Phys. Chem. B.* **2008**, *112*, 14435–14445.

(7) 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.

(8) The B3LYP methodology was recommended over MP2 methods for computations of $S_N2$reactions: Gronert, S.; Pratt, L. M.; Mogali, S. Substituent Effects in Gas-Phase Substitutions and Eliminations: β-Halo Substituents. Solvation Reverses S_N2 Substituent Effects. *J. Am. Chem. Soc.* **2001**, *123*, 3081–3091.

(9) (a) Becke, A. D. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372–1377. (b) Miehlich, B.; Savin, A.; Stoll, H.; Pruess, H. Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157*, 200–206.

(10) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(11) (a) Similar conclusions were reached by. Streitwieser, A.; Jayasree, E. G.; Leung, S. S.-H.; Choy, G. S.-C. A Theoretical Study of Substituent Effects on Allylic Ion and Ion Pair S_N2 Reactions. *J. Org. Chem.* **2005**, *70*, 8486–8491. (b) Galabov, B.; Nikolova, V.; Wilke, J. J.; Schaefer, H. J. III.; Allen, W. D. Origin of the S_N2 Benzylic Effect. *J. Am. Chem. Soc.* **2008**, *130*, 9887–9896.

(12) A similar conclusion was reached by Gronert et al.[8] in gas phase studies of substitution and elimination reactions of alkyl bromides. In another study (Gronert, S. Reactions of Gas-Phase Salts: Substitutions and Eliminations in Complexes Containing a Dianion and a Tetraalklylammonium Cation. *J. Org. Lett.* **1999**, *1*, 503−506). S_N2 barriers for methyl vs ethyl substitutions on ETMA acetate were found almost identical 25.8 vs 26.2 kcal mol^{-1} and close to those above for formate ion.

(13) Analyses of the corresponding free energy changes completely paralleled the enthalpy changes. It may be that some of the computed transition structures do not accurately portray the transition state geometry and relative energy.

(14) Uggerud, E. Nucleophilicity-Periodic Trends and Connection to Basicity. *Chem.—Eur. J.* **2006**, *12*, 1127–1136. This work dealt solely with reactions having positive activation energies and IE values between 1 and 4 eV. See also. Ochran, R. A.; Uggerud, E. S_N2 reactions with allylic substrates—Trends in reactivity. *Int. J. Mass Spectrosc.* **2007**, *265*, 169–175, for a related study with allyl systems.

(15) (a) Brauman, J. I.; Olmstead, W. N. Gas-phase nucleophilic displacement reactions. *J. Am. Chem. Soc.* **1977**, *99*, 4219–4228. (b) This may also be the case with the work of Young, L. B.; Lee-Ruff, E.; Bohme, D. K. Gas-phase nucleophilicities of the anions: H^-, F^-, OH^-, and NH_2^-. *J. Chem. Soc. Chem. Commun.* **1973**, 35, who noted that reactivities of H−, F^-, OH^-, and NH_2^- toward methyl chloride in the gas phase were approximately the same. See also Bohme, D. K.; Mackay, G. I.; Payzant, J. D. Activation energies in nucleophilic displacement reactions measured at 296 K in vacuo. *J. Am. Chem. Soc.* **1974**, *96*, 4027–4028.

(16) Li, Q.-G.; Xue, Y. Effects of Substituent and Leaving Group on the Gas-Phase S_N2 Reactions of Phenoxides with Halomethanes: A DFT Investigation. *J. Phys. Chem. A* **2009**, *113*, 10359–10366.

(17) A reviewer has pointed out that the electronic nature of single electron transfer states of the transition structures are not defined by these results.

CT900611G

JCTC Journal of Chemical Theory and Computation

Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber

Tyler Luchko,[†,¶,‡,§] Sergey Gusarov,[†] Daniel R. Roe,[||] Carlos Simmerling,[⊥,#] David A. Case,[▽,○] Jack Tuszynski,[¶,◆] and Andriy Kovalenko*[,†,‡]

National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, Alberta T6G 2M9, Canada, Department of Physics, University of Alberta, Edmonton, Alberta T6G 2G7, Canada, Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta T6G 2G8, Canada, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899-8443, Department of Chemistry, Graduate Program in Biochemistry and Structural Biology, and Center for Structural Biology, Stony Brook University, Stony Brook, New York 11794-3400, Computational Science Center, Brookhaven National Laboratory, Upton, New York 11973, BioMaPS Institute and Department of Chemistry & Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, and Department of Oncology, University of Alberta, Edmonton, Alberta T6G 2G8, Canada

Received August 18, 2009

Abstract: We present the three-dimensional molecular theory of solvation (also known as 3D-RISM) coupled with molecular dynamics (MD) simulation by contracting solvent degrees of freedom, accelerated by extrapolating solvent-induced forces and applying them in large multiple time steps (up to 20 fs) to enable simulation of large biomolecules. The method has been implemented in the Amber molecular modeling package and is illustrated here on alanine-dipeptide and protein-G.

1. Introduction

Molecular dynamics (MD) simulation with explicit solvent, in particular, available in the Amber molecular dynamics package,[1] yields accurate and detailed modeling of biomolecules (e.g., proteins and DNA) in solution, provided the processes to be described are within accessible time scales, typically up to tens of nanoseconds. A major computational burden comes from the treatment of solvent molecules (usually water, sometimes cosolvent, and counterions/buffer or salt for electrolyte solutions), which typically constitute a large part of the system. Moreover, solvent enters pockets and inner cavities of the proteins through their conformational changes, which is a very slow process and nearly as difficult to model as protein folding.

Of no surprise, then, is the considerable interest in MD simulation with solvent degrees of freedom contracted by using implicit solvation approaches. In particular, of interest is the generalized Born (GB) model,[2] in which the solvent polarization effects are represented by a cavity in dielectric continuum (optionally, with Debye screening by the charge distribution of structureless ions in the form of the Yukawa screened potential), whereas the nonelectrostatic contributions are phenomenologically parametrized against the solvent-accessible area and excluded volume of the biomolecule. The cavity shape is formed by rolling a spherical probe, of a size to be parametrized for each solvent, over the surface of the

* Corresponding author e-mail: andriy.kovalenko@nrc-cnrc.gc.ca.
† National Institute for Nanotechnology.
¶ Department of Physics, University of Alberta.
‡ Department of Mechanical Engineering, University of Alberta.
§ Present affiliation: Department of Chemistry & Chemical Biology, and BioMaPS Institute, Rutgers University, Piscataway, NJ 08854.
|| National Institute of Standards and Technology.
⊥ Stony Brook University.
Brookhaven National Laboratory.
▽ BioMaPS Institute, Rutgers University.
○ Department of Chemistry & Chemical Biology, Rutgers University.
◆ Department of Oncology, University of Alberta.

10.1021/ct900460m
Published on Web 02/01/2010

biomolecule. The polarization energy follows from the solution to the Poisson equation, which is computationally expensive, and is approximated in the GB model for fast calculation by algebraic expressions interpolating between the simple cases of two point charges in a spherical cavity. Conceptually transparent and computationally simple, the GB model has long been popular, including its implementations in the Amber molecular dynamics package.[1] However, it bears the fundamental drawbacks of implicit solvation methods: the energy contribution from solvation shell features such as hydrogen bonding can be parametrized but not represented in a transferable manner; the three-dimensional variations of the solvation structure, in particular, the second solvation shell, are lost; the volumetric properties of the solute are not well-defined; the nonelectrostatic solvation energy terms are empirically parametrized, and, therefore, effective interactions like hydrophobic interaction and hydrophobic attraction are not described from the first principles and thus are not transferable to new systems with complex compositions (e.g., with cosolvent and/or different buffer ions); and the entropic term is absent in continuum solvation, thus excluding from consideration the whole range of effects, such as the energy-entropy balance for the temperature control over supramolecular self-assembly in solution. To this end, the notion of a solvent-accessible surface, defined as that delineated by the center of the probe "rolled" over the surface, becomes meaningless for inner cavities of biomolecules hosting just a few solvent molecules.

An attractive alternative to continuum solvation is the three-dimensional molecular theory of solvation, also known as the 3D reference interaction site model (3D-RISM).[3−10] Starting from an explicit solvent model, it operates with solvent distributions rather than individual molecules, but yields the solvation structure and thermodynamics from the first principles of statistical mechanics. It properly accounts for chemical specificities of both solute and solvent molecules, such as hydrogen bonding or other association and hydrophobic forces, by yielding the 3D site density distributions of solvent, similar to explicit solvent simulations. Moreover, it readily provides via analytical expressions all of the solvation thermodynamics, including the solvation free energy potential, its energetic and entropic decomposition, and partial molar volume and compressibility. The expression for the solvation free energy (and its derivatives) in terms of integrals of the correlation functions follows from a particular approximation for the so-called closure relation used to complete the integral equation for the direct and total correlation functions.[11] The 3D-RISM theory in the so-called hypernetted chain (HNC) closure approximation was sketched by Chandler and co-workers in their derivation of density functional theory for classical site distributions of molecular liquids.[3,4] Beglov and Roux for the first time used the 3D-HNC closure to calculate the distribution of a monatomic Lennard-Jones (LJ) solvent in the neighborhood of solid substrates of arbitrary shape constructed from LJ centers[12] and introduced the 3D-RISM-HNC theory in the above way for polar molecules in liquid water.[5] Kovalenko and Hirata derived the 3D-RISM integral equation from the six-dimensional, molecular Ornstein−Zernike integral equation[11] for the solute−solvent correlation functions by averaging out the orientation degrees of freedom of solvent molecules while keeping the orientation of the solute macromolecule described at the three-dimensional level.[6,7,10] They also developed an analytical treatment of the electrostatic long-range asymptotics of both the 3D site direct correlation functions (Coulomb tails) and the total correlation functions (screened Coulomb tails and constant shifts), including analytical corrections to the 3D site correlation functions for the periodicity of the supercell used in solving the 3D-RISM integral equation.[8−10] This enabled 3D-RISM calculation of the solvation structure and thermodynamics of different ionic and polar macromolecules/supramolecules, for which distortion or loss of the long-range asymptotics for either of the correlation functions leads to huge errors in the 3D-RISM results for the solvation free energy (even for simple ions and ion pairs in water), while the analytical corrections/treatment of the asymptotics restores it to an accuracy of a small fraction of kcal/mol. Furthermore, Kovalenko and Hirata proposed the closure approximation (3D-KH closure) that couples the 3D-HNC treatment automatically applied to repulsive cores and other regions of density depletion due to repulsive interaction and steric constraints, and the 3D mean-spherical approximation (3D-MSA) applied to distribution peaks due to associative forces and other density enhancements, including long-range distribution tails for structural and phase transitions in fluids and mixtures.[7,10] The 3D-KH approximation yields solutions to the 3D-RISM equations for polyionic macromolecules, solid−liquid interfaces, and fluid systems near structural and phase transitions, for which the 3D-HNC approximation is divergent and the 3D-MSA produces nonphysical areas of negative density distributions. (For the site−site OZ, or conventional RISM theory,[11] the corresponding radial 1D-KH version is available and capable of predicting phase and structural transitions in both simple and complex associating liquids and mixtures.[10]) The 3D-RISM-KH theory has been successful in analyzing a number of chemical and biological systems in solution,[10] including structure of solid−liquid interfaces,[7] structural transitions and thermodynamics of micromicelles in alcohol−water mixtures,[13,14] structure and thermochemistry of various inorganic and (bio)organic molecules in different solvents,[15,16] conformational equilibria, tautomerization energies, and activation barriers of chemical reactions in solution,[16] solvation of carbon nanotubes,[15] structure and thermodynamics of self-assembly, stability and conformational transitions of synthetic organic supramolecules (e.g., organic rosette nanotubes in different solvents)[17−20] as well as peptides and proteins in aqueous solution,[21−23] and molecular recognition and ligand−protein docking in solution.[23,24] It constitutes a promising method to contract solvent degrees of freedom in MD simulation.

Miyata and Hirata[25] have introduced a coupling of 3D-RISM with MD in a multiple time step (MTS) algorithm, which can be formulated in terms of the RESPA[26,27] method. It converges the 3D-RISM equations for the solvent correlations at the current snapshot of the solute conformation by using the accelerated iterative MDIIS solver, then performs several MD steps, and solves the 3D-RISM

equations over again. The MDIIS (modified direct inversion in the iterative subspace) procedure[10] is a Krylov subspace type iterative solver for integral equations of liquid state theory, closely related to the DIIS approach of Pulay[28] for quantum chemistry equations and other similar algorithms, in particular, the GMRES solver.[29] The MTS approach was necessary to bring down the relatively large computational expenses of solving the 3D-RISM equations. Their implementation achieved stable simulation with the 3D-RISM equations solved at each fifth step of MD at most, which is not sufficient for realistic simulation of macromolecules and biomolecular structures of interest.

In this work, we couple the 3D-RISM solvation theory with MD in the Amber molecular dynamics package in an efficient way that includes a number of accelerating schemes. This includes several cutoffs for the interaction potentials and correlation functions, an iterative guess for the 3D-RISM solutions, and an MTS procedure with solvation forces at each MD step, which are extrapolated from the previous 3D-RISM evaluations. This coupled method makes modeling of biomolecular structures of practical interest, for example, proteins with water in inner pockets, feasible. As a preliminary illustration, we apply the method to alanine-dipeptide and protein-G in ambient water.

2. Theory and Implementation

2.1. Molecular Solvation. Solvation free energies, and their associated forces, are obtained for the solute from the 3D reference interaction site model (3D-RISM) for molecular solvation, coupled with the 3D version of the Kovalenko−Hirata (3D-KH) closure.[10] 3D-RISM provides the solvent structure in the form of a 3D site distribution function, $g_\gamma^{\mathrm{UV}}(\mathbf{r})$, for each solvent site, γ. With $g_\gamma(\mathbf{r}) \rightarrow 1$, the solvent density distribution $\rho_\gamma(\mathbf{r}) = \rho_\gamma g_\gamma(\mathbf{r})$ approaches the solvent bulk density ρ_γ. The 3D-RISM integral equation has the form:

$$h_\gamma^{\mathrm{UV}}(\mathbf{r}) = \sum_\alpha \int \mathrm{d}\mathbf{r}'\, c_\alpha^{\mathrm{UV}}(\mathbf{r}-\mathbf{r}')\chi_{\alpha\gamma}^{\mathrm{VV}}(r') \qquad (1)$$

where superscripts "U" and "V" denote the solute and solvent species, respectively; $h(\mathbf{r}) = g(\mathbf{r}) - 1$ is the site−site total correlation function; $c_\alpha^{\mathrm{UV}}(\mathbf{r})$ is the 3D direct correlation function for solvent site α having asymptotics of the interaction potential between the solute and solvent site: $c_\alpha^{\mathrm{UV}}(\mathbf{r}) \propto -u_\alpha^{\mathrm{UV}}(\mathbf{r})/(k_{\mathrm{B}}T)$; and $\chi_{\alpha\gamma}^{\mathrm{VV}}(r)$ is the site−site susceptibility of the solvent, given by

$$\chi_{\alpha\gamma}^{\mathrm{VV}}(r) = \omega_{\alpha\gamma}^{\mathrm{VV}}(r) + \rho_\alpha h_{\alpha\gamma}^{\mathrm{VV}}(r) \qquad (2)$$

Here, $\omega_{\alpha\gamma}^{\mathrm{VV}}(r)$ is the intramolecular correlation function, representing the internal geometry of the solvent molecules, while $h_{\alpha\gamma}^{\mathrm{VV}}(r)$ is the site−site radial total correlation function of the pure solvent calculated from the dielectrically consistent version of the 1D-RISM theory (DRISM).[30,31] Equation 1 is complemented with the 3D-KH closure:

$$g_\gamma^{\mathrm{UV}}(\mathbf{r}) = \begin{cases} \exp(d_\gamma^{\mathrm{UV}}(\mathbf{r})) & \text{for } d_\gamma^{\mathrm{UV}}(\mathbf{r}) \le 0 \\ 1 + d_\gamma^{\mathrm{UV}}(\mathbf{r}) & \text{for } d_\gamma^{\mathrm{UV}}(\mathbf{r}) > 0 \end{cases} \qquad (3)$$

where

$$d_\gamma^{\mathrm{UV}}(\mathbf{r}) = -\frac{u_\gamma^{\mathrm{UV}}(\mathbf{r})}{k_{\mathrm{B}}T} + h_\gamma^{\mathrm{UV}}(\mathbf{r}) - c_\gamma^{\mathrm{UV}}(\mathbf{r})$$

and $u_\gamma^{\mathrm{UV}}(\mathbf{r})$ is the 3D interaction potential of the solute acting on solvent site γ, given by the sum of the pairwise site−site potentials from all of the solute interaction sites i located at frozen positions $\mathbf{R}_i$:

$$u_\gamma^{\mathrm{UV}}(\mathbf{r}) = \sum_i u_{i\gamma}^{\mathrm{UV}}(|\mathbf{r} - \mathbf{R}_i|) \qquad (4)$$

As with the 3D-HNC closure approximation, the 3D-RISM eq 1 with 3D-KH closure 3 possesses an exact differential of the free energy and thus has a closed analytical expression for the excess chemical potential of solvation:[10]

$$\Delta\mu = k_{\mathrm{B}}T\sum_\alpha \rho_\alpha \int \mathrm{d}\mathbf{r}\Big\{\frac{1}{2}(h_\alpha^{\mathrm{UV}}(\mathbf{r}))^2\Theta(-h_\alpha^{\mathrm{UV}}(\mathbf{r})) - c_\alpha^{\mathrm{UV}}(\mathbf{r}) - \frac{1}{2}h_\alpha^{\mathrm{UV}}(\mathbf{r})c_\alpha^{\mathrm{UV}}(\mathbf{r})\Big\} \qquad (5)$$

where $\Theta(x)$ is the Heaviside function, which results in $(h_\alpha(\mathbf{r}))^2$ being applied only in areas of site density depletion.

2.2. Analytical Solvent Forces for 3D-RISM. The solvation free energy $\Delta\mu$ is generally determined by the Kirkwood "charging" formula with thermodynamic integration over the parameter λ gradually "switching on" the solute−solvent interaction potential $\tilde{u}(r;\lambda)$ along some path from no interaction at $\lambda = 0$ to the full interaction potential $u(r)$ at $\lambda = 1$. In the case of the interaction site model, it has the form:

$$\Delta\mu = k_{\mathrm{B}}T\sum_\alpha \rho_\alpha \int_0^1 \mathrm{d}\lambda \int \mathrm{d}\mathbf{r}\; g_\alpha^{\mathrm{UV}}(\mathbf{r};\lambda)\frac{\partial\tilde{u}_\alpha^{\mathrm{UV}}(\mathbf{r};\lambda)}{\partial\lambda} \qquad (6)$$

The solvation free energy $\Delta\mu(\{\mathbf{R}_i\})$ dependent on protein conformation $\{\mathbf{R}_i\}$, determined by eq 6 and obtained as eq 5, is actually the potential of mean force. The expression for the mean solvent force acting on each atom i of the solute is defined as a derivative of the solvation free energy with respect to the atom coordinates $\mathbf{R}_i$. The mean solvent force can by obtained in the general form by differentiating the expression (6) modified in such a way that the thermodynamic integration is extended over the end point $\lambda = 1$ to the full interaction potential further changed by $\mathrm{d}u_\alpha^{\mathrm{UV}}(\mathbf{r})$ due to infinitesimal shift $\mathrm{d}\mathbf{R}_i$ of solute atom i:

$$\Delta\mu(\mathbf{R}_i + \mathrm{d}\mathbf{R}_i) = k_{\mathrm{B}}T\sum_\alpha \rho_\alpha\Big(\int_0^1 \mathrm{d}\lambda \int \mathrm{d}\mathbf{r}\; g_\alpha^{\mathrm{UV}}(\mathbf{r};\lambda)\frac{\partial\tilde{u}_\alpha^{\mathrm{UV}}(\mathbf{r};\lambda)}{\partial\lambda} + \int \mathrm{d}\mathbf{r}\; g_\alpha^{\mathrm{UV}}(\mathbf{r};\lambda)\frac{\partial u_\alpha^{\mathrm{UV}}(\mathbf{r})}{\partial\mathbf{R}_i}\,\mathrm{d}\mathbf{R}_i\Big)$$

For the 3D site interaction potential (4), differentiation of this expression with respect to $\mathbf{R}_i$ immediately gives the mean solvent force acting on solute site i as

$$\mathbf{f}^{\mathrm{UV}}(\mathbf{R}_i) \equiv -\frac{\partial \Delta\mu}{\partial \mathbf{R}_i} = \sum_{\alpha} \rho_\alpha \int \mathrm{d}\mathbf{r}\, g_\alpha^{\mathrm{UV}}(\mathbf{r}) \frac{\partial u_{i\alpha}^{\mathrm{UV}}(\mathbf{r} - \mathbf{R}_i)}{\partial \mathbf{R}_i} \tag{7}$$

where $u_{i\alpha}^{\mathrm{UV}}(\mathbf{r} - \mathbf{R}_i)$ is the pairwise interaction potential between solute site i located at $\mathbf{R}_i$ and solvent site γ at $\mathbf{r}$. It is obvious that the form (7) is valid for any closure approximation that yields the solvation free energy (at a frozen solute conformation $\{\mathbf{R}_i\}$) independent of a thermodynamic integration path, that is, that possesses an exact free energy differential. These are, in particular, the 3D-HNC and 3D-KH closures.[10] The expression (7) has also been obtained, by directly differentiating a closure to the 3D-RISM equation, for the 3D-KH closure[15] and for the 3D-HNC closure.[15,25] The mean solvent force in the general form (7) still holds for any closure, subject to performing the thermodynamic integration along the path described above.

2.3. Computational Methods for Accelerating Dynamics. Modifications to the SANDER molecular dynamics module of Amber were minor. Other than calling the RISM3D subroutine, the only modifications were to add in calls for memory allocation and file input/output. A single 3D-RISM calculation is roughly 3 orders of magnitude slower than a single time step for a system solvated with the same solvent model at the same volume and density. This is not unexpected as 3D-RISM calculates the complete equilibrium distribution of solvent about the solute. To obtain meaningful sampling of solute conformations, it is necessary to reduce the computational expense of 3D-RISM calculations. To achieve this goal, three different optimization strategies were employed: (1) high-quality initial guesses to the direct correlation function were created from multiple previous solutions; (2) the pre- and postprocessing of the solute−solvent potentials, long-range asymptotics, and forces was accelerated using a cutoff scheme and minimal solvation box; and (3) direct calculation of the 3D-RISM solvation forces was avoided altogether by interpolating current force based off of atom positions from previous time steps.

2.3.1. Solution Propagation. Rapid convergence of an individual 3D-RISM calculation is facilitated by a high-quality initial guess. Given the nature of molecular dynamics simulations, it is possible to use solutions from previous time steps as the initial guess for current time step t_k. The simplest case is to use the solution from the previous time step. It is possible to improve on this by including numerically calculated derivatives:

$$c_\alpha^{\mathrm{UV}}(\mathbf{r};t_{k+1}) = c_\alpha^{\mathrm{UV}}(\mathbf{r};t_k) + (c_\alpha^{\mathrm{UV}}(\mathbf{r};t_k))' + (c_\alpha^{\mathrm{UV}}(\mathbf{r};t_k))'' + \ldots \tag{8}$$

Derivatives may be calculated for each point on the grid using finite difference techniques. In this Article, we have used up to the fourth-order derivative to calculate an initial guess:

$$c_\alpha^{\mathrm{UV},(k+1)} = c_\alpha^{\mathrm{UV},(k)} \tag{9}$$

$$c_\alpha^{\mathrm{UV},(k+1)} = 2c_\alpha^{\mathrm{UV},(k)} - c_\alpha^{\mathrm{UV},(k-1)} \tag{10}$$

$$c_\alpha^{\mathrm{UV},(k+1)} = 3(c_\alpha^{\mathrm{UV},(k)} - c_\alpha^{\mathrm{UV},(k-1)}) + c_\alpha^{\mathrm{UV},(k-2)} \tag{11}$$

$$c_\alpha^{\mathrm{UV},(k+1)} = 4c_\alpha^{\mathrm{UV},(k)} - 6c_\alpha^{\mathrm{UV},(k-1)} + 4c_\alpha^{\mathrm{UV},(k-2)} - c_\alpha^{\mathrm{UV},(k-3)} \tag{12}$$

$$c_\alpha^{\mathrm{UV},(k+1)} = 5c_\alpha^{\mathrm{UV},(k)} - 10(c_\alpha^{\mathrm{UV},(k-1)} - c_\alpha^{\mathrm{UV},(k-2)}) - 5c_\alpha^{\mathrm{UV},(k-3)} + c_\alpha^{\mathrm{UV},(k-4)} \tag{13}$$

The order at which the propagation is terminated can be indicated by the number of previous solutions used, $N^{c\mathrm{UV}}$.

2.3.2. Adaptive Solvation Box. The number of floating point entries that must be stored in memory for a 3D-RISM calculation is approximately where N_{FP} is the total number

$$N_{\mathrm{FP}} = N_{\mathrm{box}} \left[\underbrace{4}_{\text{asymptotics}} + N_{\mathrm{solv}} \left\{ \underbrace{2N_{\mathrm{MDIIS}}}_{c,\text{residual}} + \underbrace{2}_{g,h} \right\} \right] \tag{14}$$

of floating point entries, $N_{\mathrm{box}} = N_x \times N_y \times N_z$ is the total number of grid points, N_{solv} is the number of solvent atom species, and N_{MDIIS} is the number of MDIIS vectors used to accelerate convergence. A full grid for g and h is required for each solvent species, and four grids are required to compute the long-range asymptotics. Memory, therefore, scales linearly with N_{box}, while computation time scales as $O(N_{\mathrm{box}} \log(N_{\mathrm{box}}))$ due to the requirements of calculating the 3D fast Fourier transform (3D-FFT).

For independent 3D-RISM calculations, solvation box dimensions can be selected to accommodate the particular shape of the solute. For MD, however, a solvation box of fixed size throughout the simulation must be cubic to accommodate rotations and large enough to handle changes in size and shape of the solute. Alternatively, the solvation box may be determined dynamically throughout the simulation. In this case, a linear grid spacing and minimal buffer distance between any atom of the solute and the edge of the solvent box is specified. The actual dimensions of the solvent box must satisfy the constraints of maintaining specified buffer distance and linear grid spacing. To calculate the required 3D-FFT and long-range asymptotics, each grid dimension must also be divisible by 2 and have factors of only 2, 3, or 5. Previous solutions may still be propagated by transferring the past solutions to the new grid. Past solutions are truncated or padded with zeroes as required by larger or small grid dimension.

2.3.3. Potential and Force Cutoffs. Both solute−solvent potential interactions and force calculations require interactions of every solute atom with every grid point for each solvent atom species. These calculations then scale as $O(N_{\mathrm{box}} M^{\mathrm{U}} M^{\mathrm{V}})$, where M^{U} and M^{V} are the number of solute atoms and solvent species respectively. As each grid point must still be assigned a value, the use of cutoffs will not change how these calculations scale. However, computationally expensive distance-based potential calculations can be replaced with cheaper calculations outside the cutoff, reducing the computational cost by a constant factor.

As Lennard-Jones and Coulomb potentials have different long-range asymptotic behavior, the two potentials are treated

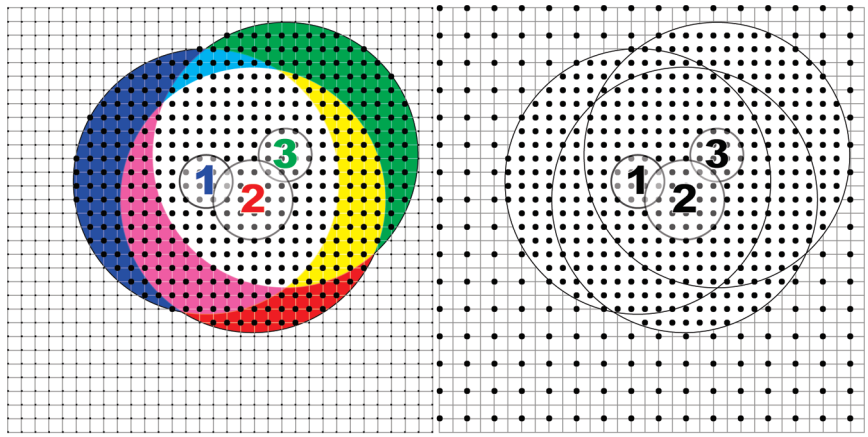


Figure 1. Cutoff schemes for grid-based (a) Lennard-Jones and (b) Coulomb potential and force calculations. Lennard-Jones calculations are performed for each solute atom only at grid sites within the cutoff distance of that atom. Grid sites within the cutoff distance of multiple solute sites take on the sum of these interactions. Coulomb interactions are calculated for every solute atom at grid sites in the union of all cutoff volumes. Grid sites outside the cutoff use explicit calculations or interpolation from surrounding values.

differently outside the cutoff radius. Lennard-Jones calculations use a hard cutoff for each solute atom. For both potential and force calculations, each solute atom only interacts with grid points within the cutoff distance, as is depicted in Figure 1a. In contrast, the long tail of the Coulomb interaction does not allow hard cutoffs to be used. Rather, within the union of the entire volume within the cutoff distance of all atoms, the entire interaction for all solute atoms is calculated at each grid point (see Figure 1b). Outside of this volume, where the interaction varies smoothly, only even grid points have the full potential calculated; that is, only one-eighth of the grid is visited. Values are then interpolated for grid points that have not been visited using a fast interpolation scheme.[32]

Contributions to atomic forces from grid points outside the cutoff volume are calculated in an analogous treatment for both Lennard-Jones and Coulomb interactions. For Lennard-Jones forces on each solute atom, only the volume within the cutoff distance from that atom is included in the integration. Coulomb forces achieve the same low density sampling used in the potential calculation by doubling the integration step size outside of the cutoff volume, effectively visiting only one-eighth of the points in this region. However, for simplicity, the cutoff volume is taken as a rectangular prism rather than a sphere for Coulomb forces alone.

An alternate method for the electrostatic potential is Ewald summation,[33] which scales as $O(N_{box}\ \ln(N_{box})M^V)$. This scaling is generally better than the cutoff method with interpolation described as $\ln(N_{box}) < M^U$ for most systems. However, the scaling coefficients for the two methods are not equal, and the cutoff method significantly outperformed Ewald summation for systems in this study. Furthermore, Ewald summation necessarily provides a periodic potential, and a correction to this must be computed to maintain the assumption of infinite dilution,[34] adding to the overhead of the Ewald method. Of course, for a large enough solute, the Ewald method with periodic correction will be more efficient than the cutoff method.

2.3.4. Force Extrapolation. A variety of multiple time step (MTS) methods have been developed to limit the number of expensive force calculations required for MD. Specifically for 3D-RISM-HNC calculations, Miyata and Hirata[25] used RESPA MTS[26,27] where slowly varying forces are only applied at an integer multiple of the base time step, effectively introducing large, periodic impulses to the dynamics. RESPA MTS has desirable properties, such as energy conservation; however, it is well-known that resonance artifacts limit the MTS step size to 5 fs for atomistic biomolecular simulations, after which the method becomes catastrophically unstable.[35−37] An alternate approach, extrapolative MTS, applies a constant force over all intermediate time steps. There are no impulses in this method to cause resonance artifacts, but it does not conserve energy as the forces at intermediate time steps do not correspond to a conservative potential. LN MTS couples extrapolative MTS with Langevin dynamics to produce stable trajectories for MTS time steps up to tens or hundreds of femtoseconds, provided the forces being extrapolated are slow varying on these time scales.[37−39] Unfortunately, the microscopic detail present in 3D-RISM calculations gives rise to forces that vary on too short a time scale to make use of LN MTS.

Inspired by LN MTS, we introduce force-coordinate extrapolation (FCE) MTS. Rather than applying a constant force, based on the last force calculations, we use previous atom configurations and forces to extrapolate what the forces should be at intermediate time steps. In this method, the forces on each of the M^U solute atoms for a current intermediate time step t_k given by the $3 \times M^U$ matrix of forces $\{\mathbf{F}\}^{(k)}$ are approximated as a linear combination of forces $\{\mathbf{F}\}^{(l)}$ at N previous time steps obtained in 3D-RISM calculations:

$$\{\mathbf{F}\}^{(k)} = \sum_{l=1}^{N} a_{kl}\{\mathbf{F}\}^{(l)}, l \in \text{ 3D-RISM steps} \qquad (15)$$

The weight coefficients a_{kl} are obtained as the best representation of the arrangement of solute atoms at the current

time step k in terms of its projections onto the "basis" of N previous solute arrangements obtained from 3D-RISM, by minimizing the norm of the difference between the current $3 \times M^{\mathrm{U}}$ matrix of coordinates $\{\mathbf{R}\}^{(k)}$ and the corresponding linear combination of the previous ones $\{\mathbf{R}\}^{(l)}$:

$$\text{minimize}|\{\mathbf{R}\}^{(k)} - \sum_{l=1}^{N} a_l\{\mathbf{R}\}^{(l)}|^2 \tag{16}$$

This is achieved by calculating the scalar products of the current coordinates matrix $\{\mathbf{R}\}^{(k)}$ and each basis coordinates matrix $\{\mathbf{R}\}^{(l)}$ and between all of the basis matrices:

$$P_{kl} = \sum_{i=1}^{M^U} (\mathbf{R}_i^{(k)} \cdot \mathbf{R}_i^{(l)}) \text{ and } S_{ll'} = \sum_{i=1}^{M^U} (\mathbf{R}_i^{(l)} \cdot \mathbf{R}_i^{(l')}) \tag{17}$$

where i is the solute atom index, and then solving the set of N linear equations for the weight coefficients a_{kl}:

$$\sum_{l'=1}^{N} S_{ll}a_{kl'} = P_{kl} \tag{18}$$

Coefficients $a_{kl'}$ are then used in eq 15 to extrapolate forces at the current intermediate time step. Similarly, the known coordinates for the current time step can be approximated from previous time steps as

$$\{\mathbf{R}\}^{(k)} = \sum_{l=1}^{N} a_{kl}\{\mathbf{R}\}^{(l)} \tag{19}$$

These forces are approximate and do not correspond to a conservative potential; thus, MD simulations using these forces will not conserve energy. However, they provide a "smooth" transition between explicitly calculated forces. As in the LN MTS method, the resulting energy gains can be damped out with the use of Langevin dynamics to provide stable, constant temperature trajectories and enhance conformational sampling through increased efficiency.[38,39]

With this method, one chooses a base time step, δt, and then calculates 3D-RISM at an integer number of base time steps, giving Δt between 3D-RISM calculations. Furthermore, RESPA MTS can also be applied to the intermediate, extrapolated forces, reducing the number of extrapolations required. As a concrete example, one can choose $\delta t = 2$ fs; after the specified number of previous coordinate sets with 3D-RISM forces has been calculated, extrapolated forces can be applied every 5 fs with new 3D-RISM solutions calculated every $\Delta t = 20$ fs.

Because the solvation forces on any particular solute atom typically correlate only with nearest neighbors, it is possible to use a cutoff for $\{\mathbf{R}\}$ and $\{\mathbf{F}\}$. Given the size of the systems in this Article, this was not used, although this capability is in our implementation.

2.3.5. Distributed Memory Parallelization. 3D-RISM calculations typically require large amounts of both computer time and memory. A distributed memory parallel implementation allows computation time to be decreased but also allows the aggregate memory of a distributed cluster to be utilized. The use of 3D-FFTs in calculating 3D-RISM solutions dictates that the memory model of the 3D-FFT library must be adopted by 3D-RISM. As we use the FFTW 2.1.5 library,[40] memory decomposition is performed along the Z-axis for all 3D arrays (u^{UV}, g^{UV}, h^{UV}, c^{UV}, etc.). Communication between processes only occurs in the MDIIS, 3D-FFT routines and for the final summation of forces.

The force extrapolation method may also be parallelized. In anticipation of the use of cutoffs, coefficients for each solute atom in eq 19 are found independently. This is trivially distributed between processes.

2.4. Solvent Model. 1D- and 3D-RISM calculate the equilibrium distribution of an explicit solvent model. Two of the most popular models for water, SPC/E[41] and TIP3P,[42] do not include van der Waals terms for the hydrogens. The incomplete intramolecular correlation in RISM theory allows a catastrophic overlap between oxygen and hydrogen sties, preventing 1D-RISM from converging on a solution. The standard approach to this problem has been to apply a small Lennard-Jones potential to the hydrogen atoms:

$$U_{\mathrm{LJ}} = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) = \varepsilon\left(\left(\frac{\sigma^*}{r}\right)^{12} - 2\left(\frac{\sigma^*}{r}\right)^{6}\right) \tag{20}$$

Common parameters used in the literature include those of Pettitt and Rossky, $\sigma = 0.4$ Å and $\varepsilon = 0.046$ kcal/mol,[43] which we will refer to as PR-SPC/E and PR-TIP3P, and those often used by Hirata and co-workers, $\sigma = 1.0$ Å and $\varepsilon = 0.05455$ kcal/mol.[44] As noted by Sato and Hirata,[45] van der Waals parameters are required to solve the RISM equations but also perturb the thermodynamics of the solution.

Alternative approaches to this problem do exist and involve corrective bridge functions[46−48] or new formalisms that go beyond RISM theory to include orientational correlations and use proper diagrams.[49−52] The major drawback of the corrective bridge function approach is that a new expression for the excess chemical potential must be derived, a nontrivial task. By including our correction in the potential, the standard closures and related thermodynamic expressions still hold. Including orientational correlations obviates the need for any "protective" Lennard-Jones potential and holds considerable promise. However, the computational complexity of these methods is even greater than that of RISM. Applying them to relatively simple systems presented here will require considerable further development of these methods.

To overcome shortcomings in previous Lennard-Jones parameters while maintaining an analytic expression for the excess chemical potential and mean solvation force, we introduce a general and transferable rule that can be applied to any model with embedded sites. Specifically, we choose

$$\frac{\sigma_{\mathrm{e}}}{2} = \frac{\sigma_{\mathrm{h}}}{2} - b_{\mathrm{he}} \tag{21}$$

$$\varepsilon_{\mathrm{e}} = 0.1\varepsilon_{\mathrm{h}} \tag{22}$$

where σ_{e} is the radius of the embedded site, σ_{h} is the radius of the host site, and b_{he} is the bond length between the two. As the embedded radius is now coincident with the host radius along the bond vector, unphysical overlap between sites is prevented. The size of ε_{e} relative to ε_{h} balances deforming the

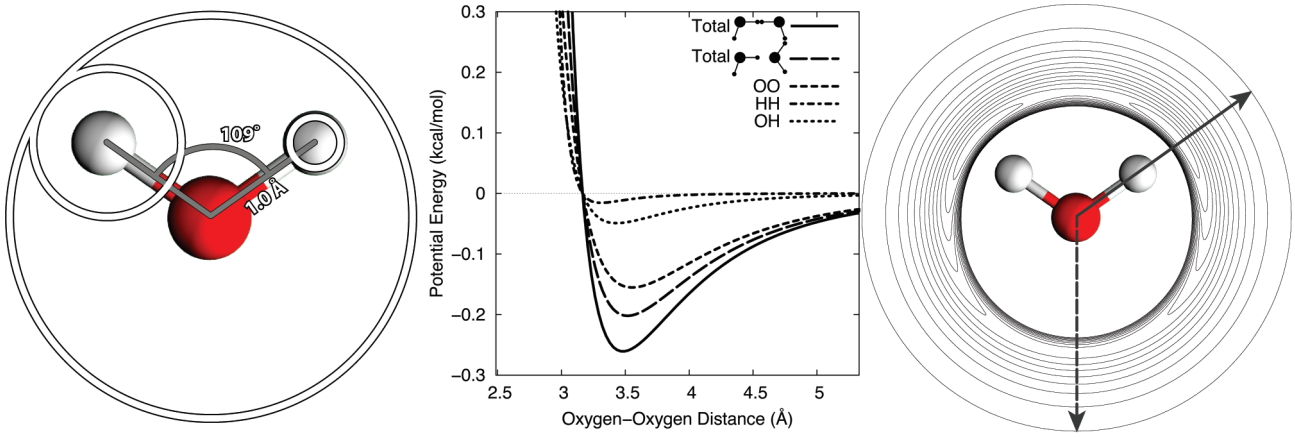


Figure 2. Modified water potential. (a) Schematic illustration of Lennard-Jones parameters for SPC/E water. Lennard-Jones radii, $\sigma/2$, are illustrated by white circles. The radius on the right-hand hydrogen corresponds to that of Pettitt and Rossky,[42] while the left-hand hydrogen radius is from eqs 21 and 22. (b) Perturbation of water−water Lennard-Jones potential due to the hydrogen potential. The maximum perturbation (solid line) is for two waters with hydrogens aligned. The case of hydrogen bonding is given by the long-dashed line, while the original potential is given by the short dashed line. HH (dot-dashed line) and OH (dotted line) interactions are a result of the new parameters. (c) Angle-dependent water−water interaction. The second water is oriented such that a hydrogen is always pointing toward the central water. The solid and long dashed arrows correspond to the solid and long-dashed lines in (b). Contour lines are spaced 0.02 kcal/mol apart.

Table 1. Parameters for Standard and Modified SPC/E and TIP3P Water Models

model name	σ_O, Å	ε_O, kcal/mol	σ_H, Å	ε_H, kcal/mol	q_O, e	q_H, e	r(OH), Å
SPC/E	3.1658	0.15530			−0.8476	0.4238	1.0000
cSPC/E	3.1658	0.15530	1.1658	0.01553	−0.8476	0.4238	1.0000
PR-SPC/E	3.1658	0.15530	0.4000	0.04600	−0.8476	0.4238	1.0000
TIP3P	3.1507	0.15200			−0.8340	0.4170	0.9572
cTIP3P	3.1507	0.15200	1.2363	0.01520	−0.8340	0.4170	0.9572
PR-TIP3P	3.1507	0.15200	0.4000	0.04600	−0.8340	0.4170	0.9572

potential of the host while proving a "stiff" enough potential to the embedded site to prevent overlaps. When applied to SPC/E and TIP3P, we refer to these models as coincident SPC/E (cSPC/E) and coincident TIP3P (cTIP3P). This is illustrated for SPC/E water in Figure 2a, and parameters for SPC/E and TIP3P water are given in Table 1.

Unlike the Pettitt and Rossky parameters, the large hydrogen site suggested here does slightly perturb the Lennard-Jones potential of the explicit model (Figure 2b and c). In particular, the well depth is increased in an orientationally dependent manner with hydrogen−hydrogen (solid line, Figure 2b) and hydrogen-bond (long dashed line) orientations becoming more favorable by 0.1 and 0.05 kcal/mol, respectively. Given the improvement in thermodynamics, this small perturbation is justified.

3. Computational Details

All simulations were carried out in a modified version of Amber 10[1] with the Langevin integrator[53] and SHAKE[54] on all bonds involving hydrogen. All 3D-RISM-KH, GB, and GBSA (GBNeck,[55] igb = 7, parameters in Amber 10) simulations for alanine-dipeptide and protein-G used free boundary conditions, no cutoff for long-range interactions, and a $\delta t = 2$ fs base time step. Explicit solvent calculations used periodic boundary conditions (PBC) with particle-mesh Ewald (PME) summation.[56]

For all alanine-dipeptide simulations, the Amber03 force field[57] was used with neutral acetyl and *N*-methal caps. Protein-G simulations used the Amber99SB force field[58] with an initial conformation from PDB ID: 1P7E.[59]

3.1. Alanine-Dipeptide − Single Point. Grid resolution and residual tolerance effects on numerical artifacts and integration of forces, including net force, were characterized with single point SPC/E 3D-RISM-KH calculations on alanine-dipeptide. A fixed solvation box of 32 Å × 32 Å × 32 Å with grid spacings of 0.5, 0.25, 0.125, and 0.0625 Å was used to perform calculations with residual error tolerances of 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. Because equilibration does not have an impact on these calculations, the default structure for alanine-dipeptide from TLEAP was used. For technical reasons, we used the Numerical Recipes FFT[60] rather than FFTW for these calculations only.

3.2. Alanine-Dipeptide − Constant Energy. Constant energy simulations were performed on alanine-dipeptide using 3D-RISM-KH and GB solvation models with the standard leapfrog-Verlet integrator. Four 3D-RISM parameter spaces were explored with 8 ns MD simulations: (1) impulse MTS 3D-RISM for a fixed box size (32 Å × 32 Å × 32 Å), using three previous solutions, with variable grid spacing (0.5 Å, 0.25 Å) and residual tolerance (10^{-3}, 10^{-4}, 10^{-5}); (2) impulse MTS 3D-RISM for a fixed box size (32 Å × 32 Å × 32 Å), 0.5 Å grid spacing and variable tolerance (10^{-3},

10^{-4}, 10^{-5}), and zero to five previous solutions ($N^{cUV} = 0...5$); (3) dynamic solvation box impulse MTS 3D-RISM calculations with buffers and cutoffs of 4, 6, 8, 10, 12, 14, 16, and 18 Å, $N^{cUV} = 5$; and (4) force extrapolation impulse MTS 3D-RISM for a fixed box size (32 Å × 32 Å × 32 Å), 0.5 Å grid spacing, 10^{-5} tolerance, and full 3D-RISM solutions every $\Delta t = 2$, 4, 6, 10, and 20 fs.

3.3. Alanine-Dipeptide − Constant Temperature. Long sampling runs were carried out on alanine-dipeptide at constant temperature (300 K) with explicit (SPC/E and TIP3P), implicit (GBNeck), and cSPC/E 3D-RISM-KH solvents. The Langevin integrator[53] was used in all cases with $\gamma = 1$ ps^{-1} for explicit solvents, $\gamma = 5$ ps^{-1} for implicit solvents, and $\gamma = 5$, 10, and 20 ps^{-1} for 3D-RISM-KH. 3D-RISM-KH simulations were performed with and without extrapolated forces. Simulations without extrapolated forces had tolerances of 10^{-5} and 10^{-3} with $N^{cUV} = 3$, and one run with a tolerance of 10^{-3} and $N^{cUV} = 5$. Simulations with force extrapolation were performed with 1 and 2 fs time steps. $\delta t = 1$ fs time step runs were performed at $\gamma = 5$, 10, and 20 ps^{-1}, used 10 previous force/coordinate pairs, and $\Delta t = 10$ or 20 fs. $\delta t = 2$ fs time step runs were performed at $\gamma = 5$, 10, and 20 ps^{-1}, used 10 previous force/coordinate pairs, and performed full 3D-RISM calculations every $\Delta t = 4$, 6, 8, or 10 fs.

For all 3D-RISM simulations, a 14 Å cutoff was used for solvent−solute potential and force calculations. Explicit solvent simulations were carried out with both 8 and 14 Å cutoffs for direct nonbond calculations. There was a negligible difference in the results, and only the 14 Å results are presented here.

All simulations were at least 3 ns. Explicit solvent simulations were extended to 21 ns to obtain better sampling. Several other simulations were extended to test convergence of sampling quality. This included GBNeck, 3D-RISM-KH with a tolerance of 10^{-3}, $N^{cUV} = 5$, and $\Delta t = 0$, and 3D-RISM-KH with $\delta t = 1$ fs, $\Delta t = 20$ fs, and $\gamma = 20$ ps^{-1}.

3.4. Sodium-Chloride. A Na^+Cl^- pair in an SPC/E solvent was simulated with 3D-RISM-KH-MD, and the distribution was compared to that expected from the potential of mean force (PMF). To prevent complete dissociation of the ion pair, a distance-based restraint was used:

$$U_{rest} = k(r - r_0)^2 \tag{23}$$

where $k = 1$ kcal/mol and $r_0 = 4$ Å. Simulations were carried out with both RESPA and FCE MTS. RESPA MTS simulations used $\Delta t = 5$ ps and $\gamma = 5$ ps^{-1}. FCE MTS simulations used $\Delta t = 10$ ps and $\gamma = 5$, 10, or 20 ps^{-1}. An integration time step of $\delta t = 1$ fs was used in all cases for a total of 500 ps simulation time.

The PMF was calculated using single point calculations of a Na^+Cl^- pair with radial separations from 2 to 8 Å in 0.02 Å steps. The expected Boltzmann probability distribution is calculated as

$$P(r)\,dr = \frac{4\pi r^2 \exp(-\beta\omega(r))\,dr}{\int_0^\infty 4\pi \exp(-\beta\omega(r))r^2\,dr} \tag{24}$$

where $\omega(r)$ is the PMF as a function of r.

3.5. Protein-G. Explicit solvent (SPC/E and TIP3P), GBSA, and cSPC/E 3D-RISM-KH simulations were carried out on protein-G (PDB ID: 1P7E).[59] SPC/E and TIP3P simulations were both solvated with 16 895 water molecules and used a 8 Å cutoff for direct, nonbonded interactions. MBondi radii were applied for the GBSA (GBNeck) system. All systems were minimized for 1000 steps. Explicit solvent systems were heated to 300 K over 10 ps before production runs. Equilibrium NPT dynamics for the explicit solvent systems was run for 3 ns. GBSA and 3D-RISM-KH were each run for 600 ps. $\gamma = 1$ ps^{-1} was used for the explicit simulations, while $\gamma = 5$ ps^{-1} was used for GBSA. 3D-RISM-KH simulations used time steps of $\delta t = 1$ fs and $\Delta t = 10$ fs. A 10 Å cutoff was used for solute−solvent calculations.

3.6. Deca-Alanine. MD, thermodynamic integration (TI), and implicit solvent free energy calculations for deca-alanine are described by Roe et al.[61] As with the implicit solvent calculations, cTIP3P 3D-RISM-KH calculations were performed on each of 1000 frames for each conformation of the 5 ns TI calculation. To accelerate the convergence of 3D-RISM solutions for each frame, the structures in each individual frame were rotated such that the first principal axis was on the z-axis using PTRAJ. A 36 Å × 36 Å × 60 Å solvation box with a 0.5 Å grid spacing was used for all calculations.

4. Results and Discussion

4.1. Decoy Analysis. Comparison of 3D-RISM-KH MD simulations to explicit and implicit solvent calculations necessarily includes the quality of the pair potential used in the 3D-RISM-KH calculation. Thus, we begin by determining our ability to reproduce the SPC/E and TIP3P model with 1D- and 3D-RISM-KH.

As all thermodynamic properties of the solvent are ultimately calculated from the 1D radial distribution function (RDF), the RDFs of our cTIP3P model with PR-TIP3P, TIP3P MD, and experimental values[62] are compared in Figure 3 (analogous SPC/E calculations show similar results). The cTIP3P parameters do not improve $g_{OO}(r)$ relative to PR-TIP3P (Figure 3a). Rather, we see that first peak has moved to a slightly larger radius, while the second peak, the so-called fingerprint of the tetrahedral hydrogen bonding of water,[43,45] is qualitatively present in PR-TIP3P but completely lost for cTIP3P. $g_{OH}(r)$ and $g_{HH}(r)$, on the other hand, are noticeably improved. The first peak of $g_{OH}(r)$ (Figure 3b) is now at the correct separation (although the magnitude is slightly too low), while the second peak is relatively unchanged. For $g_{HH}(r)$ (Figure 3c), the first peak has moved to a slightly larger separation, but the magnitude, both absolute and relative to the second peak, is much improved.

The improved structure of liquid water seen in Figure 3 should also provide improved thermodynamics, as the ultimate goal of 3D-RISM (the accurate prediction of experimental solvation free energies) is achieved through accurately reproducing the results of the explicit pair potential used as input. For the purposes of such a comparison, it is

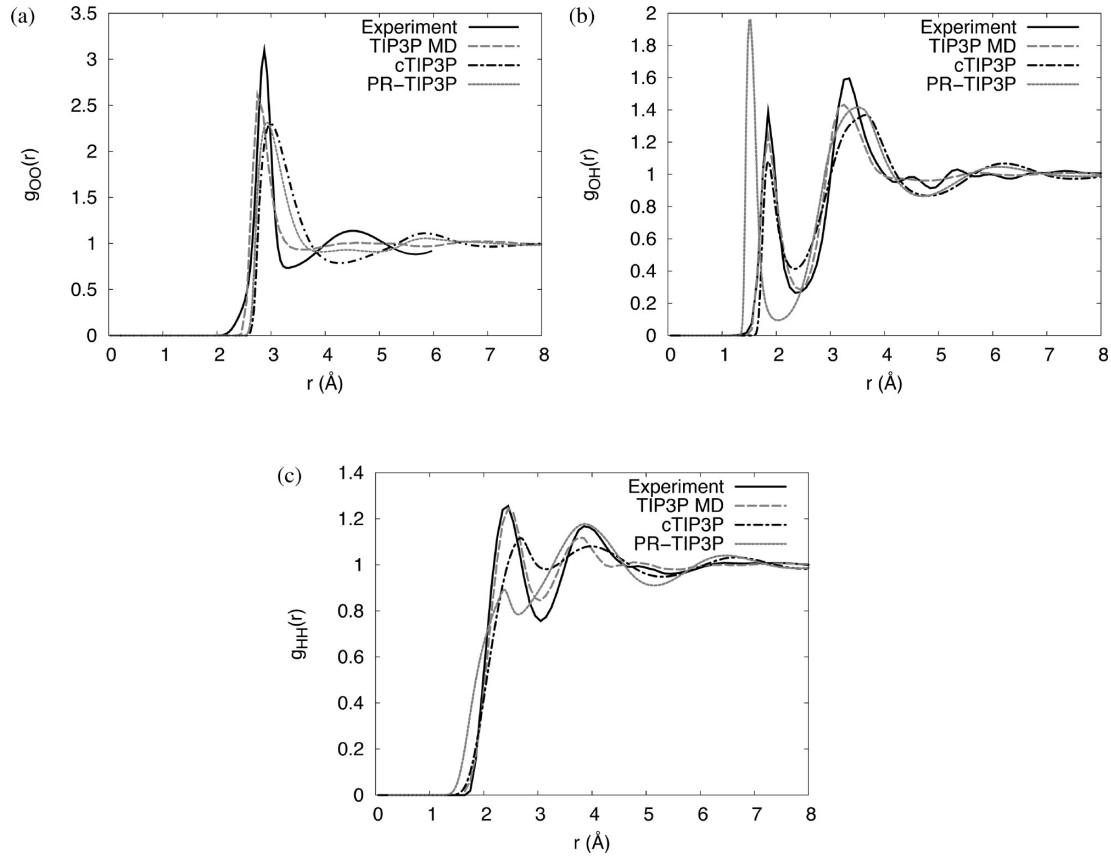


Figure 3. Water radial distribution functions from experiment, MD simulation, and 1D-RISM for (a) oxygen−oxygen,(b) oxygen−hydrogen, and (c) hydrogen−hydrogen.

useful to decompose the total solvation free energy into polar and nonpolar parts, following the standard definitions of the corresponding components in the literature:[63−65]

$$\Delta G_{\mathrm{sol}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{vdW}} + \Delta G_{\mathrm{pol}} \tag{25}$$

where G_{cav}, G_{vdW}, and G_{pol} are the free energies of cavity formation, van der Waals dispersion, and solvent polarization, respectively, and are all path-dependent quantities. 3D-RISM calculates G_{sol} directly, so to obtain each component for comparison with TI of explicit solvent it is necessary to follow the same path as used in the benchmark calculation. The free energy of solvent polarization with 3D-RISM-KH is then

$$\Delta G_{\mathrm{pol}} = \Delta G_{\mathrm{sol}} - \Delta G_{\mathrm{sol}}^{\mathrm{uncharged}} \tag{26}$$

where $G_{\mathrm{sol}}^{\mathrm{uncharged}}$ is the solvation free energy of the solute with all partial charges removed. Using this method, we can compare values for deca-alanine calculated by Roe et al.[61] (Table 2). Absolute values of solvent polarization free energy are qualitatively correct for PR-TIP3P, with alpha > left > hairpin > PP2. Both absolute values for and relative difference between the different conformations are quantitatively poor. cTIP3P greatly improves on this with relative errors of 3% or less for each conformation and less than 1 kcal/mol rmsd in relative difference between conformations. Although this does not include nonpolar contributions, it does show a good agreement with the input model.

4.2. Net Force Drift Error. A necessary property of mean solvation forces, such as those calculated by 3D-RISM, is the lack of a net force on the solute. As 3D-RISM is a grid-based method with an iterative solution, a zero net force is not guaranteed and is a function of the quality of the solution, in particular, the density of the grid and the residual tolerance of the solution. To quantify the net force error, we calculate the absolute force and root-mean-squared error (RMSE) in the force for a single point alanine-dipeptide 3D-RISM-KH solution (see Table 3).

The absolute force drift is the total force in each direction applied to the solute and should be zero for the mean solvation force. For convenience, we report the magnitude of this vector:

$$\left\| E_{\mathrm{abs}(\mathbf{f})} \right\| = \left\| \left(\sum f_x, \sum f_y, \sum f_z \right) \right\| \tag{27}$$

Ideally, all components should be zero, although in numerical force calculations this is often not the case (for example, particle-mesh Ewald summation[56]). In practice, artifacts associated with a nonzero net force can be minimized by subtracting the mass weighted average force from each atom:

$$\mathbf{f}_i' = \mathbf{f}_i - \frac{m_i}{M}\mathbf{f}_{\mathrm{net}} \tag{28}$$

where m_i is the mass of the ith solute particle and M is the total mass of the solute. However, the error in the net force is also an indicator of inaccuracies in other components not as easily corrected, such as the net torque. Table 3a suggests that the residual tolerance used for the calculation should be no higher than 10^{-3}. Values lower than this have little impact unless the grid spacing is sufficiently small.

Table 2. Comparison of Explicit TIP3P ΔG_{pol} for Deca-Alanine with 3D-RISM-KH, Poisson Equation (PE), and Generalized Born (GB)[a]

		3D-RISM[c]					
	TIP3P[b]	cTIP3P	PR-TIP3P	PE[b]	GBHCT[b]	GBOBC[b]	GBNeck[b]
				(a) ΔG_{pol}			
alpha	−44.08 ± 0.04	−44.91 ± 1.27	−55.79 ± 0.93	−47.97 ± 0.77	−51.69 ± 1.21	−49.38 ± 1.21	−43.26 ± 0.90
PP2	−76.39 ± 0.15	−76.82 ± 1.31	−93.60 ± 1.07	−78.05 ± 0.91	−77.35 ± 1.05	−78.07 ± 1.09	−77.59 ± 1.02
left	−51.30 ± 0.12	−51.60 ± 1.22	−61.81 ± 1.03	−54.85 ± 0.90	−55.05 ± 1.08	−52.67 ± 1.10	−48.19 ± 0.91
hairpin	−54.16 ± 0.25	−56.00 ± 1.17	−69.36 ± 1.31	−57.28 ± 1.13	−57.48 ± 1.45	−56.03 ± 1.47	−52.85 ± 1.29
				(b) $\Delta\Delta G_{pol}$			
PP2-alpha	−32.31	−31.91	−37.81	−30.07	−25.67	−28.69	−34.33
PP2-left	−25.09	−25.22	−31.79	−23.19	−22.31	−25.40	−29.40
PP2-hairpin	−22.23	−20.82	−24.24	−20.77	−19.87	−22.03	−24.73
alpha-left	7.22	6.69	6.02	6.88	3.36	3.29	4.93
alpha-hairpin	10.08	11.09	13.57	9.31	5.80	6.66	9.60
left-hairpin	2.86	4.40	7.55	2.43	2.43	3.37	4.67
				(c) $\Delta\Delta G_{pol}$ Root-Mean-Square Deviations			
overall		0.99	4.37	1.39	3.89	2.60	2.51
PP2		0.85	5.14	1.89	4.37	2.10	3.11
non-PP2		1.11	3.45	0.55	3.34	3.02	1.71
hairpin		1.34	3.57	1.53	2.83	2.00	1.80
nonhairpin		0.39	5.05	1.58	4.72	3.09	3.05

[a] Conformations are as in Roe et al.[60] (alpha, α-helix; PP2, polyproline II; left, left-hand helix; and hairpin, β-hairpin). Units are in kcal/mol, and errors are one standard deviation from the mean. [b] From Roe et al.[61] [c] This work.

Table 3. (a) Net Force (kcal/mol/Å), (b) Root-Mean-Squared Error in the Force, and (c) Solvation Free Energy (kcal/mol) for Single Point 3D-RISM-KH Calculations of Alanine-Dipeptide

	grid spacing			
tolerance	0.5 Å	0.25 Å	0.125 Å	0.0625 Å
		(a) Net Force		
10^{-2}	3.2	2.4	2.7	3.3
10^{-3}	1.6	0.35	0.093	0.30
10^{-4}	1.5	0.36	0.061	0.044
10^{-5}	1.5	0.37	0.041	0.0047
10^{-6}	1.5	0.37	0.042	0.0016
		(b) Force rms Error		
10^{-2}	$7.1 \times 10^{+0}$	$6.3 \times 10^{+0}$	$6.3 \times 10^{+0}$	$8.3 \times 10^{+0}$
10^{-3}	3.4×10^{-1}	1.0×10^{-1}	1.2×10^{-1}	6.2×10^{-2}
10^{-4}	1.8×10^{-1}	7.5×10^{-3}	7.6×10^{-4}	8.7×10^{-4}
10^{-5}	1.8×10^{-1}	7.4×10^{-3}	5.1×10^{-5}	9.2×10^{-6}
10^{-6}	1.8×10^{-1}	7.6×10^{-3}	5.0×10^{-5}	
		(c) Solvation Free Energy		
10^{-2}	7.5794	7.3873	7.4024	8.4253
10^{-3}	14.5614	14.4441	14.4574	14.3924
10^{-4}	14.6366	14.5097	14.5090	14.5092
10^{-5}	14.6382	14.5123	14.5121	14.5117
10^{-6}	14.6382	14.5125	14.5120	14.5116

Another method to quantify the numerical error in the forces is the RMSE.[56] For a set of "correct" forces, $\tilde{\mathbf{f}}$, we have

$$\mathrm{rmse}_f = \sqrt{\frac{\sum(\mathbf{f} - \tilde{\mathbf{f}})^2}{N_{sol}}} \tag{29}$$

Because there is no analytic calculation of the forces available for comparison, we use the solution with the smallest grid spacing (0.0625 Å) and lowest tolerance (10^{-6}) as our benchmark. As with the net force calculations, the maximum tolerance permissible is dependent on the grid spacing used. While results do improve as finer grid spacings and smaller tolerances are used, results similar to other methods, for example, particle mesh Ewald,[56] are obtained for a residual tolerance of 10^{-4} and grid spacings of 0.5 or 0.25 Å.

This observation is also evident in the solvation free energies calculated. A minimum resolution of 0.5 Å provides agreement with high grid densities within 1%. Decreasing the spacing to 0.25 Å improves this to four significant digits, but little is gained beyond this. In particular, a residual tolerance of 10^{-4} appears to be sufficient, although 10^{-3} can also be considered acceptable.

4.3. Energy Conservation. Numerical artifacts, such as those seen in the net force, typically have a large impact on energy conservation during simulation. Even after removal of the net force, all NVE simulations displayed small amplitude oscillations in the total energy about a linear decay. To quantify the linear decay, the equation

$$E_{tot} = a \cdot t + b \tag{30}$$

was fit to each data set with t representing the time in picoseconds and a corresponding to the rate of decay in kcal/mol/ps (Table 4). All calculations employed RESPA MTS, as the method is known to conserve energy for 3D-RISM time steps < = 5 fs. A comparable calculation using GBNeck yields a decay rate of $-6.37 \pm 6 \times 10^{-3}$ kcal/mol/ps.

Table 4. Rate of Decay (kcal/mol/ps) of Constant Energy Simulations of Alanine-Dipeptide for (a) Variable Grid Spacing and Solution Tolerance, (b) Variable Solution Propagation and Solution Tolerance, (c) Variable Cutoff and Solvent Box Buffer, and (d) Variable Time Step for FCE RESPA MTS[a]

(a)

	grid spacing	
tolerance	0.5 Å	0.25 Å
10^{-4}	−0.4372(9)	−0.2207(6)
10^{-5}	−0.0828(6)	−0.0824(6)
10^{-6}	−0.0234(6)	−0.0122(5)

(b)

	N^{cUV}					
tolerance	0	1	2	3	4	5
	Energy Conservation					
1×10^{-3}	0.292(3)	22.62(6)	38.6(1)	9.89(1)	0.0686(4)	0.1127(9)
1×10^{-4}	0.0063(1)	0.651(1)	0.992(4)	0.0684(2)	−0.00321(6)	0.00282(9)
1×10^{-5}	−0.00196(7)	0.01918(6)	0.01526(7)	0.00306(7)	−0.00590(6)	−0.00089(6)
	Average Number of 3D-RISM Iterations per Solution					
1×10^{-3}	47.5	18.3	21.4	28.9	29.4	35.3
1×10^{-4}	73.2	27.5	30.5	28.0	32.6	35.8
1×10^{-5}	95.7	47.1	42.8	40.4	40.1	44.1

(c)

cutoff and buffer	energy conservation
4 Å	0.623(4)
6 Å	0.0706(4)
8 Å	−0.00218(9)
10 Å	−0.00188(6)
12 Å	−0.00033(5)
14 Å	−0.00198(6)
16 Å	−0.00112(6)
18 Å	−0.00139(7)

(d)

	δt	
Δt	1 fs	2 fs
4 fs		−0.048(2)
6 fs		0.132(1)
8 fs		
10 fs	0.139(4)	
12 fs		1.15(1)
15 fs	1.50(1)	
20 fs	2.26(4)	2.37(3)
40 fs		75(3)

[a] Error in the least-squares fit for the last significant digit is given in parentheses.

The impact of grid density and residual tolerance on energy conservation is shown in Table 4a for practical grid densities. Despite differences in the net force and force RMSE produced by these two different spacings, there is negligible difference in the conservation of energy for the same residual tolerance. Considering Tables 3 and 4a suggests that the net force on the solute is primarily an artifact of the grid. The grid is part of the potential, and the tolerance determines the accuracy of the solution for this potential.

The issue is complicated by the fact that the 3D-RISM solution at each time step is not independent but is influenced by the previous solution(s) calculated and retained to seed the initial guess. Table 4b shows the effect on both energy conservation and the number of iterations required to converge on a solution for various truncations of eq 8 and residual tolerances of the 3D-RISM solution. Using zero previous solutions means that the solution at each time step is independent, $c^{UV} = 0$. A strong memory effect is observed when only one or two previous solutions are used. Increasing the number of solutions or decreasing the tolerance effectively erases this effect. Increasing the number of terms used from eq 8 increases the memory required and the number of iterations required to converge.

Two other time-saving methods introduced were cutoffs and a dynamic solvation box. In testing these methods, the cutoff was set equal to the buffer distance, effectively cutting off the corners of the solvation box. Table 4c shows that once a minimal distance of 8 Å is used, energy conservation is not affected by these methods. It should be noted that the solvation free energy calculated will vary with buffer size.

In contrast to the methods already discussed, FCE RESPA MTS (Table 4d) is not expected to conserve energy. The ability of Langevin dynamics to compensate for this depends on the rate of energy gain. For example, if the time step between 3D-RISM solutions is limited to 20 fs, energy drifts comparable to 10^{-3} tolerance are obtained. Given that dynamics are necessarily perturbed by mean-field methods like 3D-RISM and by Langevin dynamics, some energy drift may be permissible as long as the temperature and sampling are not adversely effected. Figure 4 shows the average temperature for several solvent models and parameters. Note that the values for SCP/E and TIP3P include the solute and solvent. The combination of averaging over a larger system and longer simulation time results in smaller standard errors in the mean. Combined with a sufficiently large friction coefficient, γ, a number of different parameters for FCE RESPA MTS provide stable simulations at the target temperature.

The numerical quality of the 3D-RISM solution is controlled by two parameters: (a) the residual error tolerance in the 3D-RISM calculation and (b) the linear grid spacing of the grid that the solution is found on. To a large extent,

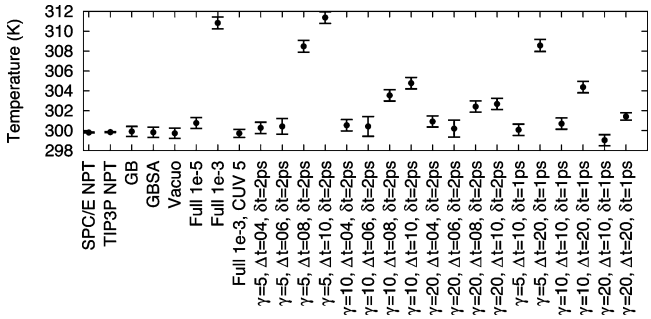


Figure 4. Average temperature for Langevin dynamics simulations of alanine-dipeptide. Error bars represent the standard error in the mean.

these two parameters independently control the conservation of energy and the net force error, respectively.

The 3D-RISM parameters used for MD/3D-RISM-KH depend on the objective of the simulation. If rigorous, constant energy simulations are desired, a residual tolerance of 10^{-5} or lower should be used with $N^{cUV} = 5$ and a buffer and cutoff of 8 Å or more. A larger buffer and cutoff, together with a finer grid spacing, provide better solvation accuracy. However, if the objective is efficient conformational sampling with solvation effects, FCE RESPA MTS can be introduced with $\Delta t = 20$ fs and a Langevin friction coefficient of $\gamma = 20$ ps^{-1}.

4.4. Sodium-Chloride. MD sampling of a Na^+Cl^- pair in solution with a weak restraint provides a simple test of the ability of FCE MTS to correctly sample a known distribution. The small size of the system (the smallest for which solvation effects will perturb the distribution) and the distance restraint near the largest potential barrier in the PMF (Figure 5a) ensure that the solvation forces play the largest possible role in the dynamics.

As expected for such a system, the FCE MTS method does cause heating that is effectively controlled by the Langevin damping coefficient. In particular, the distribution for $\gamma = 20$ ps^{-1} ((Figure 5b) is only slightly skewed from the expected distribution. Here, the distribution is shifted toward larger separations, although this is only clear by the small under sampling around global minimum.

4.5. Conformational Sampling. As 3D-RISM-KH uses an explicit solvent model as input, the conformational sampling should, ideally, be comparable to the underlying explicit solvent model used, in this case, SPC/E. Figure 6 shows free energy differences calculated from sampling distribution between SPC/E and TIP3P, GB, 3D-RISM-KH, and no model (vacuo). Figure 6a−c shows differences between other solvent models and SPC/E, providing context for comparisons with 3D-RISM-KH. Clearly, solvation effects are important, as demonstrated by Figure 6c. Even between very similar explicit models (Figure 6a), the impact can be observed with the TIP3P simulation sampling relatively more in regions of extended (−150°, 155°) and polyproline II conformations (−70°, 150°) than SPC/E. 3D-RISM-KH does see some minor deviations from the SPC/E model, with slightly more sampling of extended regions and slightly less α-helical (−58°,−47°) (Figure 6d−f). Overall, differences between 3D-RISM-KH with the cSPC/E water model and SPC/E are similar to, if slightly less than, differences between TIP3P and SPC/E. Using FCE RESPA MTS with 3D-RISM-KH also provides good results, although some softening of the potential barriers appears to occur (Figure 6f and g). This is evidenced by slightly increased sampling particularly between α-helical and polyproline II regions.

Both the quality of the sampling used for Figure 6 and the rate of convergence are shown in Figure 7. Following Lui et al.,[66] convergence of the Ramachandran sampling was calculated by dividing each trajectory into thirds and computing for each pair of trajectories, A and B:

$$(\chi^{AB})^2(t) = \frac{1}{mn}\sum_{i=1,j=1}^{m,n}(R_{i,j}^{A}(t) - R_{i,j}^{B}(t))^2 \qquad (31)$$

where the Ramachandran plot at time t is discretized into an $m \times n$ grid. The average $\chi^2(t)$ of the three trajectory combinations for each solvent model is then shown in Figure 7. As mentioned in the methods section, some trajectories were extended to obtain better sampling (explicit SPC/E and TIP3P) or to confirm that convergence was not artificial or coincidental (3D-RISM-KH with 10^{-3} tolerance and $\Delta t =$ 20 fs). As expected, the convergence rate of GBSA and 3D-RISM-KH calculations was faster than explicit solvent as friction from the solvent is removed. By this measure, 3D-RISM and GBSA sample 3−4 times more efficiently per simulation time than explicit solvent.

Electrostatic properties of the solute are strongly coupled to conformational sampling and influenced by the solvent. In particular, dielectric properties of the solvent can modify the dipole moment distribution of the solvent. The dipole moment distribution of various solvent models is shown in Figure 8 and tends to echo the results of the Ramachandran distributions. As Kwac et al.[67] have noted, peaks at 2.5, 4.5, and 7 D for alanine-dipeptide tend to correspond to extended, polyproline II, and α-helical conformations. As compared to SPC/E, all other solvent models show enhancement in extended regions and reductions in α-helical regions. Only TIP3P shows enhancement in polyproline II.

4.6. Speedup. As 3D-RISM computes the complete equilibrium solvent distribution for each solute structure it is applied to, its cost is relatively high per time step as compared to explicit solvent. To offset this, we have introduced a number of methods to reduce the number of computations required and distribute the work over multiple processors.

Serial optimizations for MD/3D-RISM-KH consist of multiple time step methods, solution propagation, cutoffs, and a dynamic solvation box. Two of these methods, MTS and solution propagation, have been previously introduced by Miyata and Hirata[25] but have been further extended here. By extending our solution propagation (eq 8) to higher derivatives, using additional previous time steps, computational efficiency has actually been slightly reduced from using only a single previous solution (Figure 9a: RESPA $N^{cUV} =$ 1 and RESPA $N^{cUV} = 5$). However, as shown in Table 4b, this additional work greatly enhances energy conservation by eliminating memory effects. A moderate speedup is still

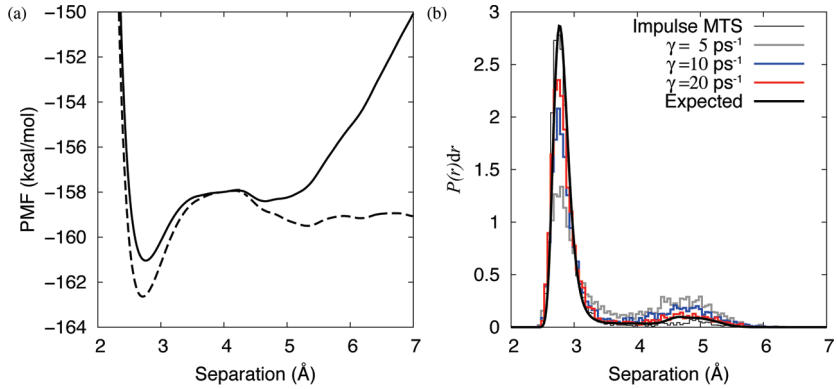


Figure 5. Na^+Cl^- pair in cSPC/E with a weak distance restraint. (a) PMF for the unrestrained pair (dash line) and restrained pair (solid line). (b) Site−site distance distribution for Na^+Cl^- with a weak harmonic restraint. The expected distribution from the potential of mean force is the thick black line; RESPA MTS is the thin black line; FCE MTS with Langevin damping coefficients of $\gamma = 5$, 10, and 20 ps^{-1} are colored gray, blue, and red, respectively.

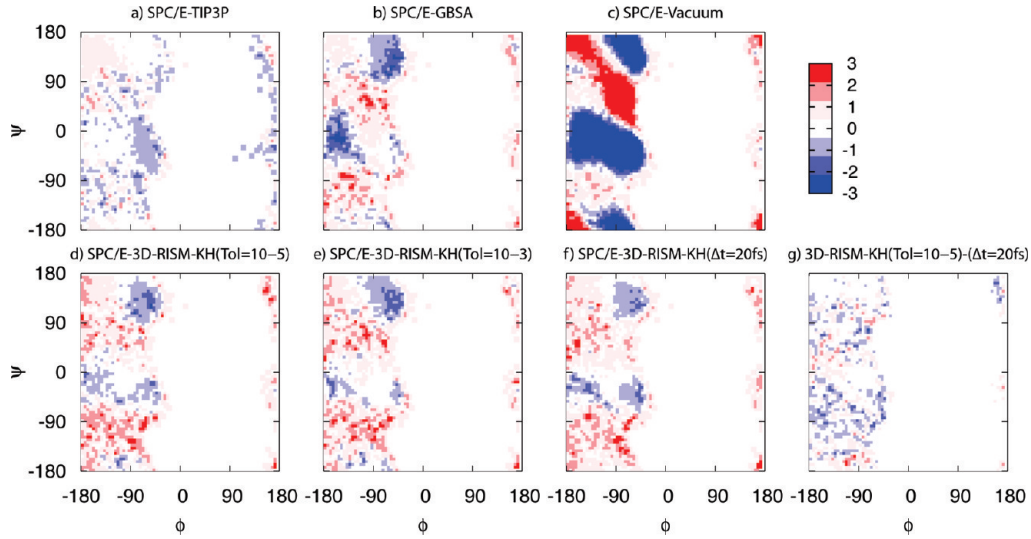


Figure 6. Ramachandran free energy differences of (a)−(f) of select solvation methods from explicit SPC/E water for alanine-dipeptide. (g) Difference of 3D-RISM-KH with a residual tolerance of 10^{-5} and 3D-RISM-KH with a FCE RESPA MTS time step of $\Delta t = 20$ fs. Energy units are in kcal/mol.

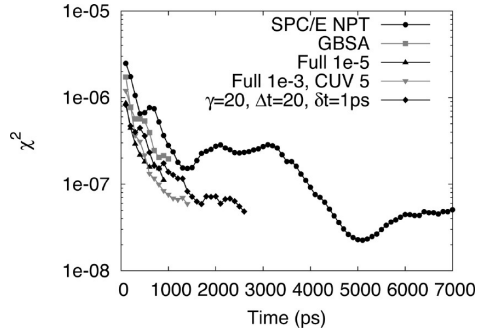


Figure 7. Convergence (χ^2) of Ramachandran plots over simulation time for select solvation methods.

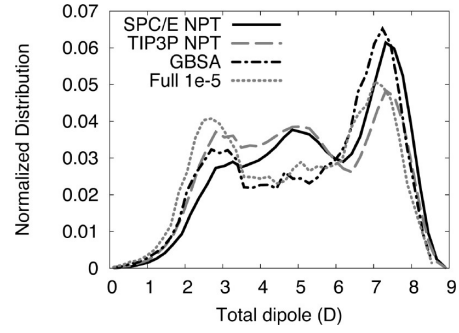


Figure 8. Dipole moment magnitude distributions of alanine-dipeptide for select solvation methods.

achieved over no solution propagation (Figure 9a: SHAKE and SHAKE $N^{cUV} = 5$).

Additional computational savings can be achieved for grid-based solute−solvent potential and force calculations. While similar to the use of cutoffs for explicit simulations, cutoffs here can take advantage of the fixed grid spacing (no need for cutoff lists), and points outside of the cutoff can still be accounted for through simple interpolation. However, cutoff methods only offer computational reductions by a constant factor as all grid points must still be visited. The computational savings are due to the number of grid points requiring expensive calculations, involving all of the solute atoms, being considerably reduced. As the grid density and number of solute atoms increase, the cutoff optimizations become more valuable.

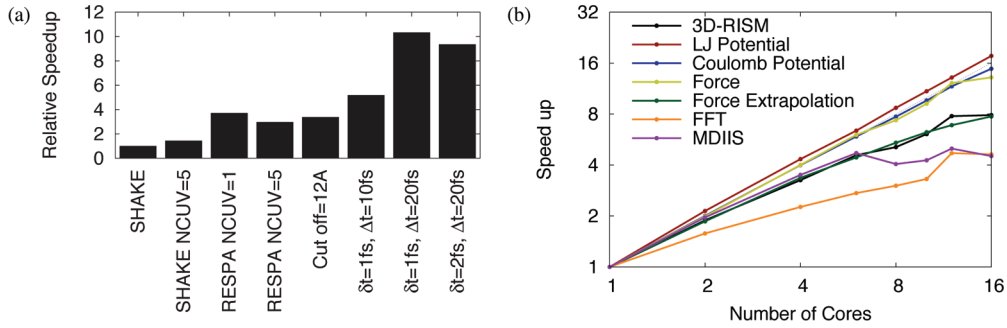


Figure 9. 3D-RISM execution speedup. (a) Serial calculations are shown with optimizations incrementally added. "SHAKE" refers to calculations where $\delta t = \Delta t$. "NCUV" indicates the number of previous solutions used for the initial guess. A cutoff of 12 Å was used for all other calculations. (b) The total parallel speedup is indicated by "3D-RISM", while the relative speedups of critical subroutines are indicated by the colored lines.

A natural extension to cutoffs is the dynamic resizing of the solvation box. For globular solutes, this has little cost-saving effect and is mostly useful as a convenience; the user only needs to input the buffer distance from the solute and the grid spacing. As solutes become less spherical or undergo large conformational changes, the benefits of the adaptive box size grow by ensuring only the minimum number of grids points is used. Together, adaptive box sizes and cutoffs offer a small overall improvement for alanine-dipeptide (Figure 9a, RESPA $N^{cUV} = 5$ and cutoff = 12 Å).

The greatest computational savings can be achieved by avoiding 3D-RISM calculations altogether by using MTS methods. The nature of biomolecular systems does not allow RESPA MTS time steps to be larger than 5 fs as resonance artifacts are introduced.[37] It is possible to overcome this resonance barrier, however, by introducing a nonconservative force approximation at intermediate time steps and using Langevin dynamics to compensate. In the case of FCE RESPA MTS, 3D-RISM-KH solutions can be calculated once every 20 fs (Figure 4). Combined with the other cost-saving measures, a speedup over a basic implementation of 3D-RISM-KH of approximately 10 times is achieved (Figure 9a, SHAKE and $\Delta t = 10, 20$ fs). While it is true that increasing the friction coefficient has a negative impact on the accuracy of dynamics, the use of a mean-field method, 3D-RISM, means that the observed dynamics are not true dynamics in any case. Our goal is to increase sampling efficiency, and using a large friction coefficient is justified in this context.

While parallelization does not decrease the computational workload, it does decrease the wall time for calculations. Furthermore, the spatial decomposition, distributed memory model used here allows the calculation to be run on a network of computers and make use of the total aggregate memory available. Relative speedups as compared to single CPU are shown in Figure 9b. Parallel speedups used protein-G simulations with a total of 50 time steps. Of these, there were eight full 3D-RISM-KH calculations, and three were interpolated 3D-RISM-KH forces. Calculations were performed on a four CPU AMD Opteron machine with four cores per CPU. Grid-based potential and force calculations that were already accelerated with cutoffs and a dynamic solvation box show linear speedups with the number of cores. The force extrapolation method has increasing efficiencies comparable

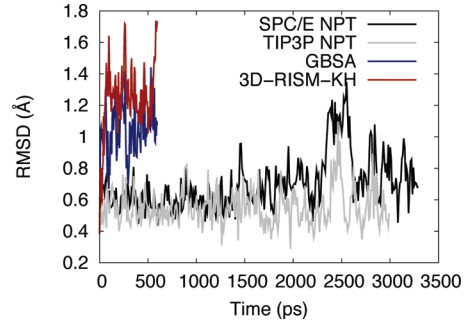


Figure 10. C_α RMSD of protein-G for explicit, implicit, and 3D-RISM-KH solvent models.

to the overall speedup of 3D-RISM. Overall parallel performance is heavily influenced by the scaling of the 3D-FFT and MDIIS routines, which also dominate the overall computation time. As we use FFTW 2.1.5 library for our 3D-FFT calculations, our speedup for the 3D-FFT part of the calculation is limited to scaling of the library.

4.7. Protein-G. Even with our decreased calculation costs, exhaustive conformational sampling of small proteins is still not accessible with 3D-RISM-KH-MD at this time. It is possible to compare different solvation models on the subnanosecond time scale for errors that may be introduced. In particular, differences in secondary and tertiary structure that are indicative of errors may be apparent in subnanosecond trajectories in 3D-RISM-KH due to the enhanced sampling that the method provides.

Figure 10 gives the root-mean-squared deviation (rmsd) of the C_α atoms from the crystal structure of protein-G as a function of simulation time. Both 3D-RISM-KH and GBSA quickly approach rmsd values of 1 Å or greater, with 3D-RISM-KH generally being higher. While these values are higher than those observed with either of the explicit models, they are comparable to previous works.[68−70] Furthermore, the RMSD of longer explicit simulations continues to grow throughout, suggesting that the equilibrium value may be close to that of 3D-RISM-KH.

Radius of gyration (Figure 11) also shows quickly equilibrating, stable trajectories for 3D-RISM-KH and GBSA with similar values and distributions. Explicit solvent simulations show a smaller and steadily increasing radius of gyration. While it is not clear if the radius of gyration has equilibrated by the end

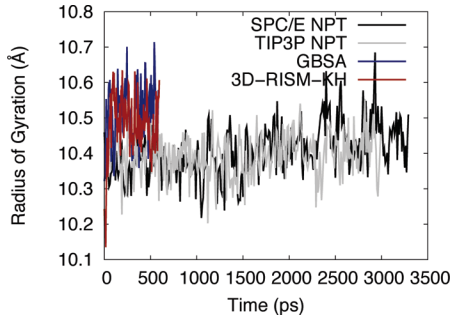


Figure 11. Radius of gyration of protein-G for explicit, implicit, and 3D-RISM-KH solvent models.

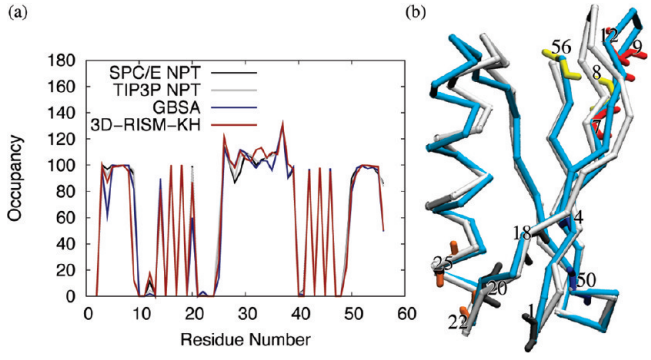


Figure 12. (a) Occupancies for internal backbone hydrogen bonding of protein-G for explicit, implicit, and 3D-RISM-KH solvent models. Occupancies >100% indicate bifurcated hydrogen bonds. (b) 3D trace of C_α atoms for NMR structure (PDB ID: 1P7E) in white and final 3D-RISM-KH structure in cyan. Backbone atoms are shown for residues with hydrogen bonding that differs from explicit solvent simulation. Images are made with VMD.[71,72]

of the simulation (3 ns), it has approached values comparable to both 3D-RISM-KH and GBSA.

As well as providing stable dynamics, solvation methods should preserve both the secondary and the tertiary structures of the solute. Hydrogen-bond calculations were performed with PTRAJ using the default criteria: a distance cutoff of 3.5 Å and an angle cutoff of 120°. Secondary structure involves hydrogen bonding within the backbone of the protein. Figure 12a shows backbone NH groups occupied by hydrogen bonds from backbone CO groups over the entire trajectory. While all solvent models are generally in good agreement, six residues show differences in the occupancies between models (Figure 12b): LYS4, GLY9, LEU12, ALA20, THR25, and GLU56. We examine these case by case.

Hydrogen bonding between residues LYS4 and LYS50 is primarily an issue for GBSA. As this is at the end of a β-sheet, it may indicate some additional flexibility, even unzipping, of the sheet. If the hydrogen-bond cutoff criteria is extended to 4.0 Å from 3.5 Å, the occupancy exceeds 80%. Enhanced flexibility also appears to be the cause for reduced hydrogen bonding between THR25 (NH) and ASP22 (CO) for nonexplicit models with 14% and 28% occupancy for GBSA and 3D-RISM-KH as compared to 40% and 50% for SPC/E and TIP3P.

The loop consisting of residues 9−12 is a site of qualitative difference in structure (Figure 12b) and behavior (Figure 13) of the 3D-RISM-KH simulation from other solvation models. While a stable hydrogen bond is seen for GBSA (81%), SPC/E (80%), and TIP3P (94%), 3D-RISM-KH shows an occupancy of only 12% for a GLY9 (NH) to LEU12 (CO). In contrast, 3D-RISM-KH also shows an occupancy of 17% for a LEU12 (NH) to GLY9 (CO), while the three other methods only show a 1−2% occupancy. This suggests that there is a oscillation between two weak hydrogen bonds. Indeed, in Figure 13b and c, the GLY9 (NH) to LEU12 (CO) hydrogen bond is disrupted by solvent. The overall effect is to bend this loop out from the protein core into the solvent (Figure 12b).

Residue ALA20 is another site where it would appear that GBSA has failed to capture the correct hydrogen bonding; however, the situation is somewhat more complex. The ALA20 (NH) site is 60% occupied in hydrogen bonding, but this bonding is strictly with THR18 (CO). For SPC/E and TIP3P, ALA20 (NH) has no hydrogen bonding with THR18, but 99% and 100%, respectively, with MET1. 3D-RISM-KH, however, has ALA20 (NH) binding to both THR18 and MET1, 54% and 33%, respectively. The correct behavior in this case is not clear. Clore and Gronenborn,[73] on the basis of nuclear magnetic resonance (NMR) data, proposed a three-site bifurcated hydrogen bond between ALA20 (NH), MET1 (CO), and a bound water molecule with residence time >1 ns. In an explicit solvent MD simulation, Sheinerman and Brooks[68] observed a long residence time water in this location, but, in this case, there was no direct hydrogen bond between ALA20 (NH) and MET1 (CO), and the water served as an intermediary between the two residues. No such long residence time water is observed in our explicit solvent simulations, although the residues are highly solvated (Figure 14a). For our 3D-RISM-KH simulation, however, the hydrogen bond is broken by the solvent (Figure 14c), reformed (Figure 14b), and broken again in the course of the 600 ps simulation.

As well, Clore and Gronenborn also proposed that a similar long residence time water would stabilize a hydrogen bond between TYR33 (NH) and ALA29 (CO). Neither the explicit simulations nor the 3D-RISM-KH simulation showed any water situated to do this, although the site was well hydrated. This is in agreement with the observations of Sheinerman and Brooks.

Both GBSA and 3D-RISM-KH have 50% occupancy for the hydrogen bond between GLU56 (NH) and ASN8 (CO) as compared to 75% and 79% for SPC/E and TIP3P. However, the reason for the low occupancy for 3D-RISM-KH is due to a larger systematic problem. As shown in Figure 15, a large, solvated cleft opens into the hydrophobic interior of the protein in the 3D-RISM-KH simulation. This allows the GLU56 (NH) and ASN8 (CO) pair to be solvated such that the hydrogen bond is disrupted. As pointed out by Kovalenko and Hirata,[48] this is likely due to the overestimation of solvent ordering around the hydrophobic side chains at the core of the protein. This is a shortcoming of the KH closure, although the same deficiency was originally identified in the HNC closure equation. As such, it is not a shortcoming of 3D-RISM and can be overcome with an improved closure, although such a development is not a trivial task.

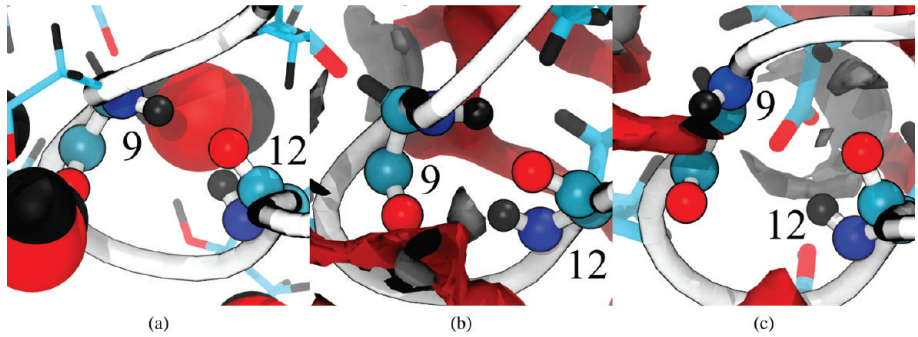


Figure 13. Backbone hydrogen bonding between residues 9 and 12 for representative structures of (a) explicit SPC/E, (b) 3D-RISM-KH with hydrogen bonding, and (c) 3D-RISM-KH without hydrogen bonding. Protein backbone drawn is as a white tube, backbone atoms for residues 9 and 12 as spheres, and side chains as sticks. Carbons are cyan, oxygens red, hydrogens black, and nitrogens blue. Solvent density isosurfaces are shown at $g_O^{UV} = g_H^{UV} = 3$ for both oxygen (red) and hydrogen (gray). Images made with VMD.[71,72]

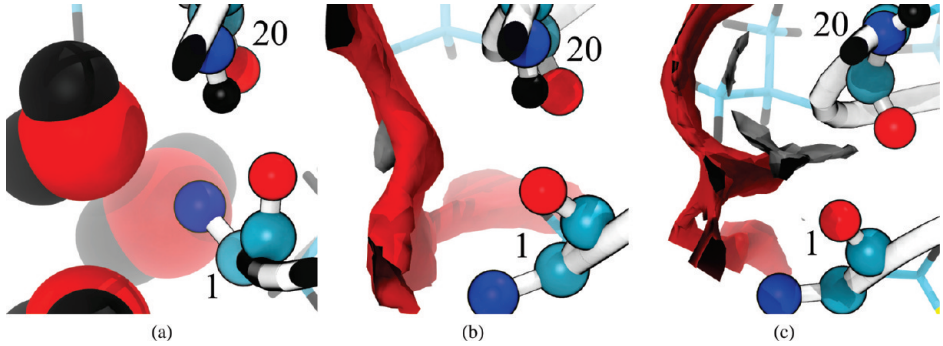


Figure 14. Backbone hydrogen bonding between residues 1 and 20 for representative structures of (a) explicit SPC/E, (b) 3D-RISM-KH with hydrogen bonding, and (c) 3D-RISM-KH without hydrogen bonding. Coloring as in Figure 13. Images made with VMD.[71,72]

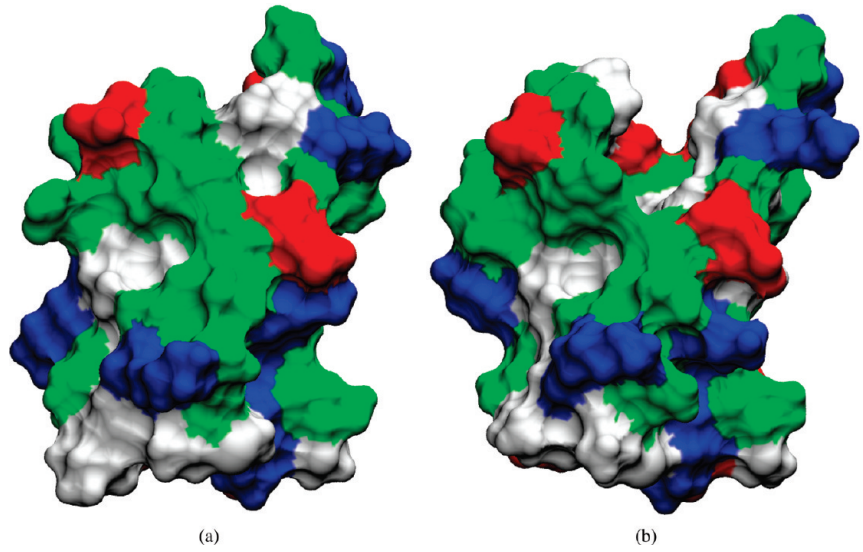


Figure 15. Solvent-accessible surface area for protein-G simulated with (a) explicit SPC/E and (b) 3D-RISM-KH. Surface is colored by residue type: acid (red), base (blue), polar (green), and nonpolar (white). Images made with VMD.[71,72,74]

5. Conclusions

We have presented an efficient coupling of molecular dynamics simulation with the three-dimensional molecular theory of solvation (3D-RISM-KH), contracting the solvent degrees of freedom, and have implemented this multiscale method in the Amber molecular dynamics package.

The 3D-RISM-KH theory uses the first principles of statistical mechanics to provide a proper account of molecular specificity of both the solute biomolecule and the solvent. This includes such effects as hydrogen bonding both between solvent molecules and between the solute and solvent, hydrophobic hydration, and hydrophobic interaction. The 3D-

RISM-KH theory readily addresses electrolyte solutions and mixtures of liquids of given composition and thermodynamic conditions. As the solvation theory works in a full statistical-mechanical ensemble, the coupled method yields solvent distributions without statistical noise, and further gives access to slow processes like hydration of inner spaces and pockets of biomolecules.

The use of 3D-RISM, a mean-field method contracting the solvent degrees of freedom in a statistical-mechanical average, means that the solvent dynamics are lost and the observed trajectories in any case are not true dynamics of MD simulation with explicit solvent. They are driven largely by a solvent-mediated potential of mean force, that is, by the probability of finding the biomolecule in a particular conformation, sampled over an ensemble of solvation shell arrangements, which frequently require extremely long time to realize (e.g., opening of protein parts to let solvent molecules or ions in to the inner spaces or pockets, multiply repeated to reach proper statistics). However, such trajectories in a solvent potential of mean force preserve the thermodynamic properties such as conformational distribution of the biomolecule and efficiently sample the conformational space regions of interest in a number of molecular biology problems such as functioning of biomolecular structures (e.g., biological channels and chaperones), protein folding, aggregation, and ligand binding.

Arrangements of solution species in the solvation shells of the biomolecule, sampled by the 3D-RISM-KH theory, can include structural solvent and/or cosolvent molecules and other associating structures like salt bridges, buffer ions, and associated ligand molecules. In the latter case, ligand molecules (or their relatively small fragments) at a given concentration in solution are described as a component of solvent at the level of site−site RISM theory and then mapped onto the biomolecule surface by the 3D-RISM method identifying the most probable binding modes of ligand molecules.[24] Together with MD sampling of biomolecular conformations, this opens up a new computational method for fragment-based drug design, which provides a proper, statistical-mechanical account of solvation forces with self-consistent coupling of both nonpolar and polar components and which gives access to binding events accompanied by rearrangements of the biomolecule and solvent on a long-time scale.

The implementation includes several procedures to maximally speed up the calculation: (i) cutoff procedures for the Lennard-Jones and electrostatic potentials and the forces acting on the solute, (ii) cutoffs and approximations for the asymptotics of the 3D site correlation functions of solvent, (iii) an iterative guess for the solution to the 3D-RISM-KH equations by extrapolating the past solutions, and (iv) multiple time step (MTS) interpolation of solvation forces between the successive 3D-RISM-KH evaluations of the forces, which are then extrapolated forward at the MD steps until the next 3D-RISM evaluation.

As a preliminary validation, we have applied the method to alanine-dipeptide and protein-G in ambient water. Analysis of the accuracy of forces, energy, and temperature, including such known artifacts as net force drift, has been performed; factors affecting the accuracy have been quantified, and the range of grid resolution and tolerance parameters ensuring reliable results has been outlined. The performance of the coupled method has been characterized and compared to MD with explicit and implicit solvent. This work is a preliminary but significant step toward the full-scale characterization and analysis of the new method and is a further improvement of its performance to address slow processes of large biomolecules in solution.

Acknowledgment. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the National Research Council (NRC) of Canada. All calculations were performed on the HPC cluster of the Center of Excellence in Integrated Nanotools (CEIN) at the University of Alberta. T.L. acknowledges financial support from the NSERC, NRC, and University of Alberta.

References

(1) Case, D.; Cheatham, T.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(2) Still, W.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(3) Chandler, D.; McCoy, J.; Singer, S. *J. Chem. Phys.* **1986**, *85*, 5971–5976.

(4) Chandler, D.; McCoy, J.; Singer, S. *J. Chem. Phys.* **1986**, *85*, 5977–5982.

(5) Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 7821–7826.

(6) Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **1998**, *290*, 237–244.

(7) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095–10112.

(8) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391–10402.

(9) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10403–10417.

(10) Kovalenko, A. In *Molecular Theory of Solvation*; Hirata, F., Ed.; Kluwer Academic Publishers: Norwell, MA, 2003; Vol. 24, Chapter 4, pp 169−276.

(11) Hansen, J.-P.; McDonald, I. *Theory of Simple Liquids*, 3rd ed.; Elsevier: Amsterdam, the Netherlands, 2006.

(12) Beglov, D.; Roux, B. *J. Chem. Phys.* **1995**, *103*, 360–364.

(13) Yoshida, K.; Yamaguchi, T.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2002**, *106*, 5042–5049.

(14) Omelyan, I.; Kovalenko, A.; Hirata, F. *J. Theor. Comput. Chem.* **2003**, *2*, 193–203.

(15) Gusarov, S.; Ziegler, T.; Kovalenko, A. *J. Phys. Chem. A* **2006**, *110*, 6083–6090.

(16) Casanova, D.; Gusarov, S.; Kovalenko, A.; Ziegler, T. *J. Chem. Theory Comput.* **2007**, *3*, 458–476.

(17) Moralez, J.; Raez, J.; Yamazaki, T.; Motkuri, R.; Kovalenko, A.; Fenniri, H. *J. Am. Chem. Soc.* **2005**, *127*, 8307–8309.

(18) Johnson, R.; Yamazaki, T.; Kovalenko, A.; Fenniri, H. *J. Am. Chem. Soc.* **2007**, *129*, 5735–5743.

(19) Tikhomirov, G.; Yamazaki, T.; Kovalenko, A.; Fenniri, H. *Langmuir* **2008**, *24*, 4447–4450.

(20) Yamazaki, T.; Fenniri, H.; Kovalenko, A. *ChemPhysChem* **2010**, *11*, 361−367.

(21) Drabik, P.; Gusarov, S.; Kovalenko, A. *Biophys. J.* **2007**, *92*, 394–403.

(22) Yamazaki, T.; Imai, T.; Hirata, F.; Kovalenko, A. *J. Phys. Chem. B* **2007**, *111*, 1206–1212.

(23) Yoshida, N.; Imai, T.; Phongphanphanee, S.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2009**, *113*, 873–886.

(24) Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A. *J. Am. Chem. Soc.* **2009**, *131*, 12430–12440.

(25) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871–882.

(26) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1991**, *94*, 6811–6815.

(27) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1992**, *97*, 1990–2001.

(28) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393–398.

(29) Saad, Y.; Schultz, M. *J. Sci. Stat. Comput.* **1986**, *7*, 856–869.

(30) Perkyns, J.; Pettitt, B. *J. Chem. Phys.* **1992**, *97*, 7656–7666.

(31) Perkyns, J.; Pettitt, B. *Chem. Phys. Lett.* **1992**, *190*, 626–630.

(32) Burkardt, J. *BLEND: Transfinite Interpolation*; http://people.scs.fsu.edu/ ~burkardt/f_src/blend/blend.html, accessed on 6/12/2008.

(33) Allen, M.; Tildesley, D. *Computer Simulation in Chemical Physics*; Oxford University Press: Oxford, UK, 1993; Chapter 4, pp 140−181.

(34) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391–10402.

(35) Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1633–1642.

(36) Barash, D.; Yang, L.; Qian, X.; Schlick, T. *J. Comput. Chem.* **2003**, *24*, 77–88.

(37) Schlick, T. *Molecular Modeling and Simulation: an Interdisciplinary Guide*, 1st ed.; Springer-Verlag New York, Inc.: Secaucus, NJ, 2002.

(38) Batcho, P.; Case, D.; Schlick, T. *J. Chem. Phys.* **2001**, *115*, 4003–4018.

(39) Qian, X.; Schlick, T. *J. Chem. Phys.* **2002**, *116*, 5971–5983.

(40) Frigo, M. *SIGPLAN Not.* **1999**, *34*, 169–180.

(41) Berendsen, H.; Grigera, J.; Straatsma, T. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(42) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926–935.

(43) Pettitt, B.; Rossky, P. *J. Chem. Phys.* **1982**, *77*, 1451–1457.

(44) Hirata, F.; Levy, R. M. *Chem. Phys. Lett.* **1987**, *136*, 267–273.

(45) Sato, H.; Hirata, F. *J. Chem. Phys.* **1999**, *111*, 8545–8555.

(46) Cortis, C.; Rossky, P.; Friesner, R. *J. Chem. Phys.* **1997**, *107*, 6400–6414.

(47) Du, Q.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **2000**, *104*, 796–805.

(48) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *113*, 2793–2805.

(49) Dyer, K.; Perkyns, J.; Stell, G.; Pettitt, B. *J. Chem. Phys.* **2008**, *129*, 104512–104520.

(50) Gendre, L.; Ramirez, R.; Borgis, D. *Chem. Phys. Lett.* **2009**, *474*, 366–370.

(51) Sumi, T.; Sekino, H. *J. Chem. Phys.* **2006**, *125*, 034509−034509.

(52) Urbic, T.; Vlachy, V.; Kalyuzhnyi, Y.; Dill, K. *J. Chem. Phys.* **2003**, *118*, 5516–5525.

(53) Loncharich, R.; Brooks, B.; Pastor, R. *Biopolymers* **1992**, *32*, 523–535.

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

(55) Mongan, J.; Simmerling, C.; McCammon, J.; Case, D.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

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

(57) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(58) Simmerling, C.; Strockbine, B.; Roitberg, A. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.

(59) Ulmer, T.; Ramirez, B.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179–9191.

(60) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in FORTRAN: the Art of Scientific Computing*, 2nd ed.; Cambridge University Press: New York, NY, 1992.

(61) Roe, D.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(62) Soper, A.; Phillips, M. *Chem. Phys.* **1986**, *107*, 47–60.

(63) Still, W.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(64) Levy, R.; Zhang, L.; Gallicchio, E.; Felts, A. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.

(65) Gonçalves, P.; Stassen, H. *Pure Appl. Chem.* **2004**, *76*, 231–240.

(66) Liu, P.; Kim, B.; Friesner, R.; Berne, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749–13754.

(67) Kwac, K.; Lee, K.-K.; Han, J. B.; Oh, K.-I.; Cho, M. *J. Chem. Phys.* **2008**, *128*, 105106.

(68) Sheinerman, F.; C. B. III. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 193–202.

(69) Patel, S.; A. M. Jr.; C. B. III. *J. Comput. Chem.* **2004**, *25*, 1504–1514.

(70) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins: Struct., Funct., Genet.* **2001**, *45*, 144–158.

(71) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(72) Stone, J. M. Sc. thesis, Computer Science Department, University of Missouri-Rolla, 1998.

(73) Clore, G.; Gronenborn, A. *J. Mol. Biol.* **1992**, *223*, 853−856.

(74) Sanner, M.; Olsen, A.; Spehner, J.-C. *Fast and Robust Computation of Molecular Surfaces. Proceedings of the 11th ACM Symposium on Computational Geometry*, New York, 1995; pp C6−C7.

CT900460M

JCTC Journal of Chemical Theory and Computation

Entropy from Correlations in TIP4P Water

Emanuela Giuffré,[†] Santi Prestipino,[†] Franz Saija,[‡] A. Marco Saitta,[§] and Paolo V. Giaquinta*,[†]

Università degli Studi di Messina, Dipartimento di Fisica, Contrada Papardo, 98166 Messina, Italy, CNR − Istituto per i Processi Chimico−Fisici, Sede di Messina, Contrada Papardo, Viale Ferdinando Stagno d'Alcontres 37, 98158 Messina, Italy, and Université Pierre et Marie Curie − Paris 06, UMR 7590, IMPMC, F-75015 Paris, France

Received November 24, 2009

Abstract: We use molecular dynamics to compute the pair distribution function of liquid TIP4P water as a function of the intermolecular distance and of the five angles that are needed to specify the relative position and orientation of two water molecules. We also calculate the translational and orientational contributions to the two-body term in the multiparticle correlation expansion of the configurational entropy at three selected thermodynamic states, where we also test various approximations for the angular dependence of the pair distribution function. We finally compare the results obtained for the pair entropy of TIP4P water with the experimental values of the excess entropy of ordinary water.

I. Introduction

It is superfluous to motivate the persisting interest in a deeper understanding and more effective modeling of the equilibrium structure of water.[1] The paradigmatic status of this molecular liquid in the realms of natural and life sciences, at the interface between a variety of disciplines such as physics, chemistry, and biology, justifies the continuing efforts that are being made to refine experimental tools and theoretical approximations so as to achieve more and more reliable predictions of the microscopic and macroscopic properties of this substance as well as of the many anomalous aspects that mark its thermodynamical, structural, and dynamical behavior in a unique way.[2]

The main object of the present study is the calculation of the pair entropy of liquid water, i.e., the contribution of two-body density correlations to the configurational entropy. The statistical-mechanical framework is that provided by the multiparticle correlation expansion of the excess entropy, originally derived for a classical atomic fluid in the canonical ensemble[3,4] and later extended to the grand-canonical ensemble,[5,6] the two expressions having in fact been shown to be formally equivalent.[7] Many other authors have discussed this topic; we refer the reader to ref 8 for a short commented list of some relevant contributions to the subject. To our knowledge, the first calculation of the pair entropy of a pure molecular fluid dates back to the seminal paper authored by Lazaridis and Karplus,[9] who investigated the interplay between orientational correlations and entropy in the TIP4P model of liquid water.[10] The TIP4P potential is an effective pairwise four-site potential: a Lennard-Jones interaction site is located on the oxygen, two positive charges on the hydrogens, and an extra negative charge is located away from the oxygen along the hydrogen−hydrogen bisector. Intramolecular degrees of freedom are neglected in the TIP4P model, but nonetheless, this potential turns out to be a good transferable potential in that it ultimately provides a qualitatively correct representation of the phase diagram of water, also in comparison with other interaction models, except at very high pressure.[11,12]

The calculation of the pair entropy for a given model potential requires the knowledge of the pair distribution function (PDF) of the molecular liquid, a quantity that, in the case of TIP4P water, depends on six variables, viz., the distance between the centers of mass of the two molecules and five angles that are needed to specify their relative

* Corresponding author e-mail: paolo.giaquinta@unime.it.
[†] Università degli Studi di Messina.
[‡] CNR − Istituto per i Processi Chimico−Fisici.
[§] Université Pierre et Marie Curie.

10.1021/ct900627q
Published on Web 02/22/2010

orientation. Such a function cannot be easily obtained from numerical simulation since a statistically reliable result requires a massive numerical sampling to be carried out over a discrete six-dimensional grid whose spacing is consequently the outcome of a delicate compromise between the available computer power and the desired quantitative accuracy of the calculation. This is precisely the reason why Lazaridis and Karplus evaluated the pair entropy of liquid TIP4P water using a variety of approximations for the five-dimensional orientational distribution function (ODF) which, computed at a given intermolecular separation and then multiplied by the ordinary radial distribution function (RDF), eventually yields the full six-dimensional PDF. Such approximations implemented partial representations of the ODF on the basis of combinations of the monovariate and bivariate angular distribution functions as obtained from numerical simulation.

A few other attempts have as yet followed the route traced by Lazaridis and Karplus for the calculation of the pair entropy of water. Giaquinta and co-workers exploited one of the approximate schemes illustrated in ref 9, the so-called "adjusted gas-phase" (AGP) approximation, to highlight different ordering regimes in liquid TIP4P water[13] and to "measure" the relative amounts of positional and angular order through the translational and orientational pair entropies, used for the first time together as unbiased, self-consistent, and independent "order parameters" to map the phase diagram of water.[14] The first "exact" calculation—i.e., one performed without resorting to any approximate partial representation of the ODF—of the pair entropy was carried out by Zielkiewcz in four different models of computer water at ambient conditions.[15,16] A temperature analysis of the results for the single-point charge (SPC) model was later presented by the same author.[17] More recently, Wang and co-workers suggested a novel nonparametric approach to computing the pair entropy, alternative to the histogram-based method, and further based on a generalized Kirkwood superposition approximation (GKSA) for the ODF.[18] This approach was tested in five water models at ambient conditions.

In this paper, we present the results of a numeric calculation of the *full* PDF of liquid TIP4P water, carried out with the method of molecular dynamics (MD), at ambient conditions as well as in two other thermodynamic states, at lower temperature and higher pressure, respectively. The general theoretical framework is introduced and discussed in section II, together with the approximations that were implemented for the ODF in addition to the direct MD calculations, whose technical details are summarized in section III. The resulting translational and orientational pair entropies are presented in section IV and therein compared with the experimental data for the excess entropy of ordinary water. Section V is finally devoted to concluding remarks.

II. Theoretical Framework

A. Pair Entropy of a Molecular Liquid. Statistical mechanics provides a general expression for the entropy of a classical, atomic or molecular, fluid which, in general, can be written as an infinite sum of contributions associated with spatially integrated n-point density correlations:

$$S_{\mathrm{ex}} = \sum_{n=2}^{\infty} S_n \tag{1}$$

where S_{ex} is the excess (with respect to the corresponding ideal gas) entropy. In the absence of external fields, the two-body term—that, in the following, we shall refer to as the "pair entropy"—ordinarily delivers the dominant contribution to the excess entropy of a liquid.[13,19,20] As such, S_2 has been often used as an approximate "local" (in thermodynamic space) estimate of S_{ex} in that the calculation of the pair entropy does not require a thermodynamic potential to be integrated along an extended path connecting the state whose entropy one is interested in with another (reference) state where the thermodynamic properties of the fluid are known by independent means, as in the high-temperature ideal-gas asymptotic regime, or can be computed with other techniques.[12,21]

The pair entropy of molecular fluids reads:[9,22]

$$S_2 = -\frac{1}{2}k_{\mathrm{B}}\left(\frac{\rho}{\Omega}\right)^2 \int [g(\mathbf{r}_1, \mathbf{r}_2, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \ln g(\mathbf{r}_1, \mathbf{r}_2, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) - g(\mathbf{r}_1, \mathbf{r}_2, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) + 1]\, \mathrm{d}\mathbf{r}_1\, \mathrm{d}\mathbf{r}_2\, \mathrm{d}\boldsymbol{\xi}_1\, \mathrm{d}\boldsymbol{\xi}_2 \tag{2}$$

where k_{B} is the Boltzmann constant, ρ is the particle number density, and $g(\mathbf{r}_1, \mathbf{r}_2, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ is the PDF which, in general, depends on the vector radii $(\mathbf{r}_1, \mathbf{r}_2)$ of the molecular centers of mass and on the pair of Euler angles sets $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$, where $\boldsymbol{\xi}_\alpha \equiv \{\theta_\alpha, \phi_\alpha, \chi_\alpha\}$ specifies the absolute orientation of the αth molecule in the laboratory reference frame $\{\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}\}$. The three Euler angles are respectively defined in the ranges $0 \le \theta_\alpha \le \pi$, $0 \le \phi_\alpha < 2\pi$, and $0 \le \chi_\alpha < 2\pi$, with angular elements $\mathrm{d}\boldsymbol{\xi}_\alpha = \sin(\theta_\alpha)\, \mathrm{d}\theta_\alpha\, \mathrm{d}\phi_\alpha\, \mathrm{d}\chi_\alpha$. Correspondingly, $\Omega \equiv \int \mathrm{d}\boldsymbol{\xi}_\alpha = 8\pi^2$. We note that the set of variables introduced above is in fact redundant for a homogeneous and isotropic molecular fluid since, in the absence of external fields, the PDF depends on the *relative* position and orientation of the two molecules. For water molecules, this information can be encoded into six variables only, viz., the radial separation r between the centers of mass plus five angles. In fact, let us label the oxygen and hydrogen atoms of the αth molecule as O_α, $H_{1\alpha}$, and $H_{2\alpha}$, respectively; let us further identify the mean point of the segment $H_{1\alpha}H_{2\alpha}$, joining the two hydrogen atoms, as D_α. In terms of the unit vectors

$$\hat{\mathbf{z}} = \frac{\overrightarrow{O_1O_2}}{|\overrightarrow{O_1O_2}|}, \qquad \hat{\mathbf{z}}_\alpha = \frac{\overrightarrow{O_\alpha D_\alpha}}{|\overrightarrow{O_\alpha D_\alpha}|} \tag{3}$$

one has for the angle θ_α formed by the dipole moment of the αth molecule, lying in the direction $\hat{\mathbf{z}}_\alpha$, and the intermolecular axis $\hat{\mathbf{z}}$:

$$\theta_\alpha = \arccos(\hat{\mathbf{z}}_\alpha \cdot \hat{\mathbf{z}}) \ (0 \le \theta_\alpha \le \pi) \tag{4}$$

In order to define the other Euler angles, we need to specify a reference system $\{\hat{\mathbf{x}}_\alpha, \hat{\mathbf{y}}_\alpha, \hat{\mathbf{z}}_\alpha\}$ sticking out of each molecule. This can be done by defining, for instance, the $\hat{\mathbf{x}}_\alpha$ and $\hat{\mathbf{y}}_\alpha$ axes as

$$\hat{\mathbf{x}}_\alpha = \frac{\overrightarrow{H_{2\alpha}H_{1\alpha}}}{|\overrightarrow{H_{2\alpha}H_{1\alpha}}|}, \qquad \hat{\mathbf{y}}_\alpha = \hat{\mathbf{z}}_\alpha \wedge \hat{\mathbf{x}}_\alpha \tag{5}$$

Hence, in terms of the auxiliary unit vector

$$\hat{\mathbf{N}}_\alpha \equiv \frac{\hat{\mathbf{z}} \wedge \hat{\mathbf{z}}_\alpha}{|\hat{\mathbf{z}} \wedge \hat{\mathbf{z}}_\alpha|} \tag{6}$$

specifying the line of nodes that is associated with the αth molecule, we obtain

$$\phi_\alpha = \arg(\hat{\mathbf{x}} \cdot \hat{\mathbf{N}}_\alpha, \hat{\mathbf{y}} \cdot \hat{\mathbf{N}}_\alpha) \ (0 \le \phi_\alpha < 2\pi) \tag{7}$$

and

$$\chi_\alpha = \arg(\hat{\mathbf{x}}_\alpha \cdot \hat{\mathbf{N}}_\alpha, -\hat{\mathbf{y}}_\alpha \cdot \hat{\mathbf{N}}_\alpha) \ (0 \le \chi_\alpha < 2\pi) \tag{8}$$

where $\arg(u, v)$ is the polar argument of the vector (u, v). The angle χ_α describes the rotation of the αth water molecule around its own dipole vector. One can readily see that one of the above six angles is superfluous; in fact, upon choosing, say, $\hat{\mathbf{x}} \equiv \hat{\mathbf{N}}_1$, the angle ϕ_1 identically vanishes. In the following, we shall refer to this set of angular coordinates as the *A* set. Actually, one may further take advantage of a symmetry of the PDF that follows from the observation that the pointing directions of the unit vectors $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ depend on the (arbitrary) way one has labeled the two hydrogen atoms within each water molecule. In turn, this choice uniquely determines the pointing directions of $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$. This residual freedom reflects on the behavior of the (exact) PDF that is in fact invariant under χ_α rotations moving this angle to $[\chi_\alpha + \pi(\mathrm{mod}\ 2\pi)]$. Hence, it should suffice to consider the dependence of the PDF on χ_α within the interval $[0, \pi]$. We shall refer to this modified—by effect of symmetry—set of variables as the A_s set.

The above choice of variables closely recalls the one made by Lazaridis and Karplus.[9] However, in addition to the pairs (θ_1, θ_2) and (χ_1, χ_2), they used the angle $\phi \equiv \phi_1 - \phi_2$ describing the relative rotation of a pair of molecules (more specifically, of their respective line-of-nodes unit vectors, $\hat{\mathbf{N}}_1$ and $\hat{\mathbf{N}}_2$) around the intermolecular axis. Moreover, their conventions on how to measure the two pairs of angles mentioned above differ from those specified for the set *A*. In particular, H_{11} was explicitly taken to be the hydrogen atom of molecule 1 that is closest to the oxygen atom of molecule 2, and similarly for H_{21} (see Figure 1 of ref 9 and the relative caption). Lazaridis and Karplus further exploited a number of symmetries of the PDF involving all five angles. In fact, in addition to those introduced above which concern the pair (χ_1, χ_2), they also considered two residual symmetries of the PDF: they noted that the two water molecules are interchangeable, which allows one to integrate the angle θ_2 from θ_1 to π, and also observed that the angle ϕ can be integrated from 0 to π only. We shall refer to this set of angular coordinates as the B_s set which, however, we shall implement without resorting to the additional symmetry concerning the angle θ_2.

A still different choice (*C*) was made by Zielkiewicz,[15,16] who assumed $\{\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}\} \equiv \{\hat{\mathbf{x}}_1, \hat{\mathbf{y}}_1, \hat{\mathbf{z}}_1\}$, the orthonormal triad attached to molecule 1 being specified as above by eqs 3 and 5. He then considered the spherical coordinates of O_2 in the reference frame centered on O_1:

$$\Theta = \arccos(\hat{\mathbf{r}} \cdot \hat{\mathbf{z}}_1), \qquad \Phi = \arg(\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{r}}_\perp, \hat{\mathbf{y}}_1 \cdot \hat{\mathbf{r}}_\perp) \tag{9}$$

where $\mathbf{r} \equiv \overrightarrow{O_1O_2}$ and $\mathbf{r}_\perp = \mathbf{r} - (\mathbf{r} \cdot \hat{\mathbf{z}}_1)\,\hat{\mathbf{z}}_1$, with $0 \le \Theta \le \pi$ and $0 \le \Phi < 2\pi$, in addition to the three Euler angles $(\theta_2, \phi_2, \chi_2)$ that specify the global orientation of the second molecule with respect to the central one, according to eqs 4, 7, and 8. As observed before, the alternative for $\hat{\mathbf{x}}_\alpha$ is 2-fold, allowing one to restrict the range of ϕ_2 and χ_2 to the interval $[0, \pi]$. The set of variables where this latter symmetry property is employed will be referred to as the C_s set.

It is important to realize that different choices of the angles that specify the relative orientation of two water molecules, as well as of the associated intervals within which the PDF is then being sampled, lead to different histograms and, consequently, to potentially different numerical estimates of the pair entropy. This is obviously due to the finite integration meshes as well as to the number of symmetries that are implemented in each angular set. In fact, we should expect that the best results will be achieved when one makes full use of the symmetries of the PDF, which is actually the case of the B_s set. In this respect, we note that, for an assigned number of configurations over which the PDF histogram is being sampled, the information contents of the histograms built by using the A_s and C_s sets should be four times larger than that of the corresponding histograms in the *A* and *C* sets. On the other hand, the statistical quality achieved with the B_s set should be twice as good as that achieved with the A_s and C_s sets.

Now, let $\mathbf{\Psi}$ be any 5-fold set of independent angular variables. Upon observing that $\int \mathrm{d}\boldsymbol{\xi}_1\, \mathrm{d}\boldsymbol{\xi}_2 = 2\pi \int \mathrm{d}\mathbf{\Psi}$, we can rewrite eq 2 as

$$s_2 = -k_\mathrm{B}\left(\frac{2\pi}{\Omega}\right)^2 \rho \int [g(r, \mathbf{\Psi}) \ln g(r, \mathbf{\Psi}) - g(r, \mathbf{\Psi}) + 1] r^2\, \mathrm{d}r\, \mathrm{d}\mathbf{\Psi} \tag{10}$$

where s_2 is the pair entropy per particle. Following Lazaridis and Karplus,[9] we factorize the PDF as

$$g(r, \mathbf{\Psi}) = g(r) g(\mathbf{\Psi}|r) \tag{11}$$

where $g(r)$ is the RDF and $g(\mathbf{\Psi}|r)$ is the conditional distribution for the relative orientation of two molecules separated by a distance *r*, i.e., the quantity that we shall refer to in the following as the ODF. The following normalization condition holds:

$$\frac{2\pi}{\Omega^2} \int g(r, \mathbf{\Psi})\, \mathrm{d}\mathbf{\Psi} = g(r) \tag{12}$$

Correspondingly, the pair entropy can be resolved into the sum of two terms:

$$s_2 = s_2^{(\mathrm{tr})} + s_2^{(\mathrm{or})} \tag{13}$$

where

$$s_2^{(\mathrm{tr})} = -k_\mathrm{B}(2\pi)\rho \int_0^\infty [g(r) \ln g(r) - g(r) + 1] r^2\, \mathrm{d}r \tag{14}$$

is the translational pair entropy and

$$s_2^{(or)} = \int_0^\infty \rho\, g(r) S^{(or)}(r) r^2 dr \quad (15)$$

is the orientational pair entropy that is defined in terms of the orientational local entropy (OLE):

$$S^{(or)}(r) = -k_B \left(\frac{2\pi}{\Omega}\right)^2 \int g(\mathbf{\Psi}|r) \ln g(\mathbf{\Psi}|r)\, d\mathbf{\Psi} \quad (16)$$

B. Approximations for the Pair Distribution Function. At variance with the RDF, the numerical evaluation of the ODF is a formidable computational task because of the number of variables this function depends on. Approximating the ODF may be in many cases the only viable route to the calculation of the orientational contribution to the thermodynamic properties of liquid water. In this respect, Lazaridis and Karplus proposed a number of approximate schemes for the full PDF, essentially based on the assumption that the ODFs of gaseous and liquid water have similar short-range structures.[9] The first "family" of such approximations was generated by factorizing the ODF into a product of one-dimensional (1d) and two-dimensional (2d) marginals, defined as the probability distributions of one angle, or the joint probability distributions of two angles, regardless of the values attained by the remaining angles:

$$g(\Psi_i|r) \equiv \frac{\int g(\mathbf{\Psi}|r)\, d^4\Psi_{j\neq i}}{\int d^4\Psi_{j\neq i}} \quad (17)$$

and

$$g(\Psi_i, \Psi_j|r) \equiv \frac{\int g(\mathbf{\Psi}|r)\, d^3\Psi_{k\neq i,j}}{\int d^3\Psi_{k\neq i,j}} \quad (18)$$

Lazaridis and Karplus (LK) introduced and discussed various factorizations that were tested against the thermodynamic properties (energy, entropy) of the gas phase of TIP4P water.[9] They eventually concluded that, overall, the most balanced scheme was the one which they referred to as the F7 factorization:

$$g_{LK}^{(F7)}(\theta_1, \theta_2, \phi, \chi_1, \chi_2) \equiv \left[\frac{g(\theta_1, \theta_2)\, g(\chi_1, \chi_2)\, g(\theta_1, \chi_2)\, g(\theta_2, \chi_1)}{g(\theta_1)\, g(\theta_2)\, g(\chi_1)\, g(\chi_2)}\right] g(\phi) \quad (19)$$

where the parametric dependence on the radial separation r is implicit in both the full and marginal ODFs that appear on the left- and right-hand sides of eq 19, respectively. This approximation exploits the "flatness" of the ODF with respect to the angle ϕ and the ensuing absence, in the liquid phase, of significant correlations between ϕ and the other angles.[9]

In addition to the factorization scheme illustrated above, Lazaridis and Karplus described a different approach based on the low-density limit of the ODF, modified by a subset of the marginal ODFs that were obtained from the numerical simulations of *liquid* TIP4P water. The resulting AGP approximation for the ODF can be written as follows:

$$g_{LK}^{(AGP)}(\theta_1, \theta_2, \phi, \chi_1, \chi_2) = g_s(\theta_1, \theta_2, \phi, \chi_1, \chi_2) \cdot C_{LK}(\theta_1, \theta_2, \phi, \chi_1, \chi_2) \quad (20)$$

where

$$C_{LK}(\theta_1, \theta_2, \phi, \chi_1, \chi_2) = \frac{g(\theta_1, \theta_2)}{g_s(\theta_1, \theta_2)} \cdot \frac{g(\chi_1, \chi_2)}{g_s(\chi_1, \chi_2)} \cdot \frac{g(\phi)}{g_s(\phi)} \quad (21)$$

and where we have again suppressed the dependence on the intermolecular distance r. In eq 21, $g(\theta_1, \theta_2)$, $g(\chi_1, \chi_2)$, and $g(\phi)$ are the "exact" marginals, which can be calculated by numerical simulation, while the function g_s and the associated marginals follow from

$$g_s(\mathbf{\Psi}|r) \equiv g_{gas}(\mathbf{\Psi}|r) + \mathcal{J}(r)[1 - g_{gas}(\mathbf{\Psi}|r)] \quad (22)$$

where

$$g_{gas}(\mathbf{\Psi}|r) = \left(\frac{\Omega^2}{2\pi}\right) \frac{\exp[-\beta u(r, \mathbf{\Psi})]}{\int \exp[-\beta u(r, \mathbf{\Psi})]\, d\mathbf{\Psi}} \quad (23)$$

with $u(r, \mathbf{\Psi})$ being the interaction potential and

$$\mathcal{J}(r) = \frac{U_{MD}(r) - U_{gas}(r)}{U(r) - U_{gas}(r)} \quad (24)$$

In eq 24, $U_{MD}(r)$ is the angularly averaged interaction energy evaluated at a specified distance r through the MD simulation, $U_{gas}(r)$ is the corresponding quantity calculated in the gas phase, and $U(r)$ is the unweighted average of $u(r, \mathbf{\Psi})$ over orientations. The ODF that follows from eq 20 must be eventually normalized so as to satisfy eq 12. The smoothing function $\mathcal{J}(r)$ is such that the angularly averaged energy evaluated through the function g_s reproduces the quantity $U_{MD}(r)$.

As seen from eqs 20 and 21, the naive gas-phase approximation—formulated in eqs 22, 23, and 24—is *adjusted* so as to enforce, through the C_{LK} factor, a reasonable degree of consistency between the resulting ODF and a subset of liquid-phase marginal distributions. Wang and co-workers[18] recently noted that a non-negligible correlation exists between the angles θ and χ of each water molecule which cannot be accounted for by the mere product of the corresponding 1d marginals, i.e., $g(\theta)$ and $g(\chi)$. Hence, they proposed a GKSA factorization of the ODF that differs from the F7 scheme set up by Lazaridis and Karplus in that the modified ODF includes an extra pair of 2d "intramolecular" marginals, viz., $g(\theta_1, \chi_1)$ and $g(\theta_2, \chi_2)$. Motivated by this observation, we also report in this paper on three variants of the original AGP approximation where the coupling factor C_{LK}, which appears in eq 20, has been replaced, in turn, by the following expressions:

$$C_1(\theta_1, \theta_2, \phi, \chi_1, \chi_2) = C_{LK}(\theta_1, \theta_2, \phi, \chi_1, \chi_2) \left[\frac{g(\theta_1, \chi_1)}{g_s(\theta_1, \chi_1)} \cdot \frac{g(\theta_2, \chi_2)}{g_s(\theta_2, \chi_2)}\right] \quad (25)$$

$$C_2(\theta_1,\theta_2,\phi,\chi_1,\chi_2) = C_{LK}(\theta_1,\theta_2,\phi,\chi_1,\chi_2) \times \left[\frac{g(\theta_1,\chi_2)}{g_s(\theta_1,\chi_2)} \cdot \frac{g(\theta_2,\chi_1)}{g_s(\theta_2,\chi_1)}\right] \quad (26)$$

$$C_3(\theta_1,\theta_2,\phi,\chi_1,\chi_2) = C_2(\theta_1,\theta_2,\phi,\chi_1,\chi_2) \times \left[\frac{g(\theta_1,\chi_1)}{g_s(\theta_1,\chi_1)} \cdot \frac{g(\theta_2,\chi_2)}{g_s(\theta_2,\chi_2)}\right] \quad (27)$$

We shall refer to the three modified AGP schemes reported in eqs 25, 26, and 27 as the AGP1, AGP2, and AGP3 approximations, respectively. We observe that the AGP2 approximation includes the same 2d marginals that also appear in the F7 approximation originally proposed by Lazaridis and Karplus,[9] while the AGP3 coupling factor reproduces the factorization of 2d marginals adopted by Wang and co-workers.[18]

In order to simplify the calculation of the pair entropy, Lazaridis and Karplus further resorted to an approximate computational strategy. In fact, they averaged the liquid-phase angular distributions that appear in both the F7 and AGP schemes over three distinct spatial regions ("shells") whose boundaries were chosen so as to coincide with the positions of the first peak and of the first two troughs in the oxygen−oxygen RDF of TIP4P water at 25 °C and 1 atm. Their choice leads to the following intervals: $0 \le r \le 0.28$ nm, $0.28 \text{ nm} < r \le 0.34$ nm, and $0.34 \text{ nm} < r \le 0.56$ nm. As for the three gas-phase marginals that also contribute to the AGP approximation, they were apparently calculated at three "representative" distances corresponding to the three regions over which the same marginals were calculated in the liquid by simulation, i.e, 0.28 nm, 0.32 nm, and 0.45 nm. We note that a similar computational strategy was adopted by Wang and co-workers[18] who also computed the OLE over three consecutive shells that closely correspond to the analogous choice made by Lazaridis and Karplus. However, they used a different approach to estimate the pair entropy, viz., the so-called *k*th nearest-neighbor method.

The AGP approximation clearly depends on the preknowledge of the RDF spatial profile of the model in a given thermodynamic state. We propose a generalization of this method by averaging the liquid-phase marginals over a variable number of shells, $N_{shell} = R_{max}/\Delta R_{shell}$, where R_{max} is the chosen distance cutoff and ΔR_{shell} is the width of each shell. Obviously, the maximum number of shells over which the marginals can be computed corresponds to a discretization of the interval such that $\Delta R_{shell} = \Delta r$, where Δr is the spatial resolution of the calculation. In addition, the gas-phase marginals are computed at the midpoint of each interval. We shall refer to this modified implementation of the AGP approximation as the multishell AGP (MSAGP) approximation.

III. Simulation Method and Technical Details

We carried out constant-pressure constant-temperature MD simulations of the TIP4P model, implemented through the PINY code.[23] The system contained 512 molecules in a cubic cell with periodic boundary conditions. The spherical cutoff of the Lennard-Jones interactions was 0.8 nm. Electrostatic interactions were modeled with the particle-mesh Ewald method. A time step of 2.5 fs turned out to be sufficient for a proper dynamical evolution since the TIP4P model is based on a rigid-molecule description. Typical simulation times were in the 50 ns range, corresponding to runs 2×10^7 steps long. The MD configurations were stored every 0.5 ps in such a way that the PDF was calculated over as many as 10^5 system snapshots. Under ambient conditions, the molecular density was initially 1 g cm^{-3}, corresponding to a width of the simulation cell of 2.48 nm.

As already emphasized, the evaluation of the pair entropy from eq 10 is far from being a trivial task, it being the outcome of a six-dimensional integration. In practice, many aspects of the calculation may seriously influence the final numerical accuracy such as (i) the number, N_{conf}, of configurations contributing to the thermal averages, a parameter that affects the "quality" of the integrand, i.e., its regularity as a function of the independent variables; (ii) the integration mesh sizes (Δr, $\Delta\mathbf{\Psi}$); (iii) the numerical integration method, whose impact on the result is not *a priori* obvious, especially when the integrand happens to be noisy.

The RDF and the full PDF were respectively estimated through the following formulas:

$$g(r) = I(r;\Delta r)\left\{\frac{4\pi}{3}\rho\left[\left(r+\frac{\Delta r}{2}\right)^3 - \left(r-\frac{\Delta r}{2}\right)^3\right]\right\}^{-1} \quad (28)$$

$$g(r,\mathbf{\Psi}) = I(r,\mathbf{\Psi};\Delta r,\Delta\mathbf{\Psi})\left\{\frac{4\pi}{3}\rho\left[\left(r+\frac{\Delta r}{2}\right)^3 - \left(r-\frac{\Delta r}{2}\right)^3\right]\right\}^{-1}\left(\frac{2\pi}{\Omega^2}\Delta\mathbf{\Psi}\right)^{-1} \quad (29)$$

where $I(r;\Delta r)$ and $I(r,\mathbf{\Psi};\Delta r,\Delta\mathbf{\Psi})$ are the corresponding histograms and

$$\Delta\mathbf{\Psi} = \left[-\cos\left(\theta_1+\frac{\Delta\theta_1}{2}\right)+\cos\left(\theta_1-\frac{\Delta\theta_1}{2}\right)\right]\times \left[-\cos\left(\theta_2+\frac{\Delta\theta_2}{2}\right)+\cos\left(\theta_2-\frac{\Delta\theta_2}{2}\right)\right]\times \Delta\chi_1\Delta\chi_2\Delta\phi_2 \quad (30)$$

We also took advantage of the symmetry properties of $I(r,\mathbf{\Psi};\Delta r,\Delta\mathbf{\Psi})$, with the effect of doubling the statistics whenever the range of a given angle is halved from 2π to π. We assumed $\Delta r = 0.01$ nm and $\Delta\Psi_i = 10°$ for each of the five angles that were involved in the calculation. This latter value turned out to be a reasonable compromise between histogram resolution and statistics, and moreover, it is the value that has been commonly used in the past literature on the subject.[9,15,18] All the integrations were performed using the standard Simpson method. In order to get a quantitative feeling on the numerical error associated with the choices made for Δr and $\Delta\Psi_i$, we compared the RDF obtained directly from the simulation with the output from the normalization condition reported in eq 12 in all the thermodynamic states investigated. We found that the relative deviation was about 3% at very short distances, in the region corresponding to the initial rise of the RDF from zero up to its highest maximum, and rapidly dropped to zero

Table 1. Translational and Orientational Pair Entropies (e.u.) of TIP4P Water at 298 K[a]

source	$s_2^{(tr)}$	$s_2^{(or)}$	s_2
Lazaridis and Karplus[9]	−3.14	−9.1[b]	−12.2
Lazaridis and Karplus[9]	−3.14	−11.7[c]	−14.8
Wang et al.[18]	−3.15	−10.52	−13.67
Zielkiewicz[15,16]	−2.97[d]	−11.9[e]	−14.9[f]

[a] Data from refs 9 and 18 refer to constant-pressure simulations carried out at 1 atm, while data from refs 15 and 16 refer to constant-volume simulations corresponding to a density ρ_m = 0.999 g cm^{-3}. [b] Estimate obtained using the F7 approximation. [c] Estimate obtained using the AGP approximation. [d] The estimate originally provided in ref 15 has been corrected by adding the missing contribution reported in ref 16. [e] Estimate inferred upon subtracting the translational pair entropy from the cumulative pair entropy. [f] Estimate inferred upon subtracting the ideal-gas entropy from the absolute entropy reported in ref 16.

with increasing distances, already being less than 0.1% for $r \geq 0.3$ nm.

IV. Results

We investigated the properties of TIP4P water in three thermodynamic states located in the stable liquid-phase region, i.e., (T = 260 K, P = 1 bar), (T = 300 K, P = 1 bar), and (T = 300 K, P = 4 kbar). We shall first present the results obtained at ambient conditions, which we shall compare with those obtained by other authors using the Monte Carlo or the molecular dynamics method, either resorting to approximate representations of the ODF[9,18] or carrying out a full direct calculation.[15,16] We shall then illustrate how the pair entropy changes upon lowering the temperature down to 260 K or increasing the pressure up to 4 kbar.

A. TIP4P Water at Ambient Conditions. We computed equilibrium averages at ambient conditions (T = 300 K, P = 1 bar) over sets of up to 10^5 MD configurations. The resulting average values of the specific density ρ_m and of the excess internal energy U_{ex} were 0.986 g cm^{-3} and −10.01 kcal mol^{-1}, respectively. Under such thermodynamic conditions, the cumulative ideal-gas contribution to the entropy of TIP4P water, modeled as a gas of noninteracting rigid molecules, amounts to 30.74 entropy units (e.u.; 1 e.u. = 4.184 J K^{-1} mol^{-1}),[24] about one-third of which (10.44 e.u.) is to be ascribed to the rotational degrees of freedom. We report in Table 1 the currently available estimates of the translational, orientational, and cumulative pair entropies of TIP4P water at ambient conditions for a comparison with the present results.

The translational pair entropy can be computed in a straightforward way since it requires the RDF only as an input (see Figure 1). Upon using eq 14, we obtained −3.05 e.u. for this quantity.

1. MD Results for the Orientational Pair Entropy. The OLE computed directly from simulation—i.e., using no approximate partial representation—is shown in Figure 2 for all the angular sets introduced in section II.B. The discrepancies observed between some of the five estimates are almost entirely due to their different statistical qualities, as already discussed in section IIB. This aspect is clarified in Figure 3 where we reported the OLEs obtained using the

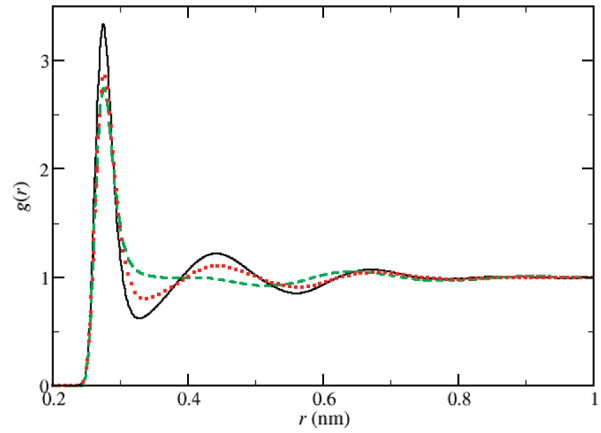


Figure 1. Radial distribution function of TIP4P water averaged over 10^5 configurations: (T = 260 K, P = 1 bar), black continuous curve; (T = 300 K, P = 1 bar), red dotted curve; (T = 300 K, P = 4 kbar), green dashed curve.

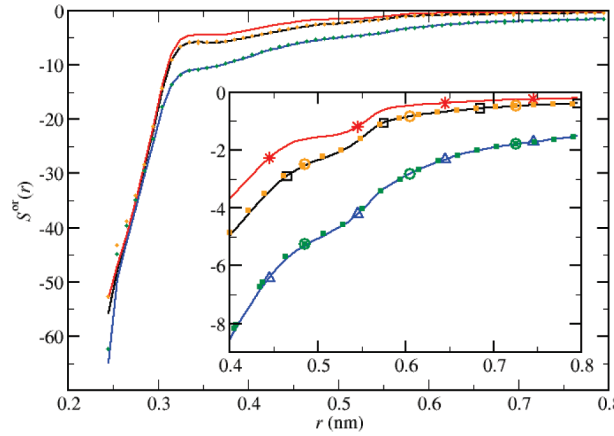


Figure 2. Orientational local entropy (e.u.) of TIP4P water at ambient conditions averaged over 10^5 MD configurations: *A* set, blue solid curve (further marked with triangles in the inset); *C* set, green dotted curve (further marked with circles in the inset); A_s set, black solid curve (further marked with squares in the inset); C_s set, orange dotted curve (further marked with circles in the inset); B_s set, red solid curve (further marked with asterisks in the inset).

(A, C), (A_s, C_s), and B_s angular sets, the corresponding histograms now being averaged over sets of configurations whose numbers lie in the ratio 8:2:1, respectively, as suggested by the number of symmetries employed in each angular set. As a result, the five estimates are manifestly seen to collapse onto the same curve. This comparison confirms that the B_s set provides the most accurate estimate that can be generated with an assigned number of configurations.

Figure 4 shows the function $I_S^{(or)}(r) = \rho\, g(r)\, S^{(or)}(r)\, r^2$, that yields upon integration the orientational pair entropy (see eq 15), depicted for increasing values of N_{conf}. The most relevant feature of this function is the profound minimum located at r = 2.75 nm, i.e., where the RDF attains its maximum value corresponding to the first coordination shell. We observe that the shape of the minimum, aside from its depth, does not appreciably change with the number of configurations; on the other hand, the longer-range part of the function shifts almost rigidly toward zero as N_{conf} increases. We found that, even upon sampling the $I_S^{(or)}(r)$

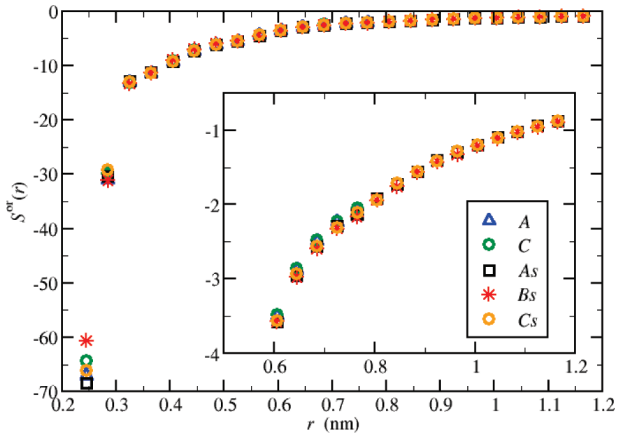


Figure 3. Orientational local entropy (e.u.) of TIP4P water at ambient conditions averaged over 8×10^4 configurations for the *A* and *C* sets, over 2×10^4 configurations for the A_s and C_s sets, and over 1×10^4 configurations for the B_s set.

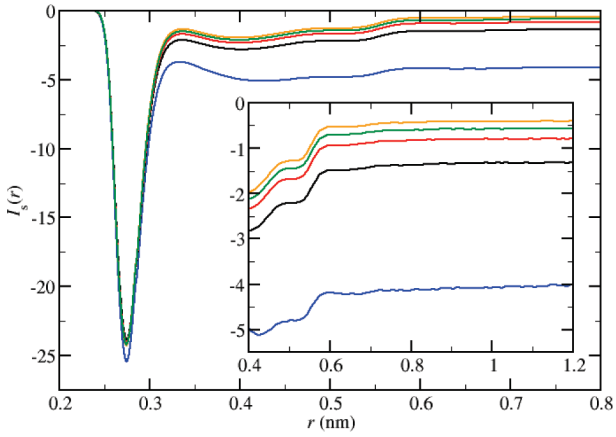


Figure 4. The function $I_S^{(or)}(r) = \rho\, g(r)\, S^{(or)}(r)\, r^2$, computed for the B_s angular set at ambient conditions, plotted as a function of the distance for increasing values of the number of configurations $N_{conf} = \{1, 3, 5, 7, 10\} \times 10^4$. The inset shows the long-range behavior of the function, decreasing in absolute terms with increasing N_{conf} at fixed *r*.

histogram over 10^5 configurations, the function has not decayed yet to zero over distances corresponding to half the width of the simulation cell. This behavior suggests that, over such intermolecular separations, the decorrelation time of the angular degrees of freedom is on the order of the time (~50 ns) spanned by the longest MD trajectory generated in the present calculations. The concurring effect of the finite mesh size cannot be excluded as well. Upon integrating the most refined histogram produced for $I_S^{(or)}(r)$, we obtained −14.7 e.u. for the orientational pair entropy, a value that is certainly underestimated because of the arguments illustrated before. Hence, in order to obtain a more reliable estimate, we resorted to an extrapolation of the quantity $\sigma(r; N_{conf}) \equiv g(r; N_{conf})\, S^{(or)}(r; N_{conf})$, that was modeled, as a function of N_{conf}, according to the following inverse-power law:

$$\sigma(r;N_{conf}) = \tilde{\sigma}(r) + \frac{A(r)}{N_{conf}^{\alpha(r)}} \qquad (31)$$

where $\tilde{\sigma}(r)$, $A(r)$, and $\alpha(r)$ are *r*-dependent parameters that were determined through a least-squares fit of the MD data.

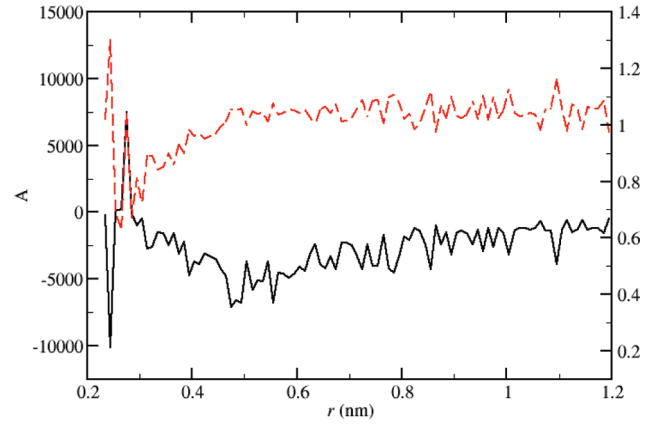


Figure 5. Space dependent amplitude (continuous black curve, left axis) and exponent (broken red curve, right axis) of the extrapolating function used for $\sigma(r, N_{conf})$ in the B_s set at ambient conditions.

This procedure differs in a significant way from the one used in ref 15, where, instead, a series of estimates of the absolute entropy, obtained over MD trajectories of increasing length, was extrapolated as a function of the simulation time. However, as noted above, such estimates are likely affected by a truncation error arising from the nonvanishing tail of the OLE. This is the reason why we extrapolated this function, and only afterward performed the integration in eq 15.

The current fit was carried out over seven sets of data corresponding to values of N_{conf} ranging between 4×10^4 and 10×10^4, with sequential increments of 10^4 configurations. Moreover, every set was assigned a weight proportional to the number of configurations used to calculate the function $\sigma(r; N_{conf})$. The resulting best-fit values of the parameters $A(r)$ and $\alpha(r)$ are plotted in Figure 5 as a function of *r*. Notwithstanding their apparently noisy aspect, the quality of the fit was good everywhere: in fact, the minimum root-mean-square deviation from the MD data turned out to be less than 10^{-5} e.u. in the region of the first minimum—where, as noted before, the OLE does not exhibit a marked dependence on N_{conf} —decreasing further with distance by more than 2 orders of magnitude for $r > 0.6$ nm. The long-range trend of $\alpha(r)$ indicates that the tail of $\sigma(r; N_{conf})$ deviates from $\tilde{\sigma}(r)$ approximately as N_{conf}^{-1}. The asymptotic estimate of the integrand function, $\tilde{I}_S^{(or)}(r)$, that we obtained for the B_s set is depicted in Figure 6. Upon comparing this result with the present largest N_{conf} estimate of the same quantity, one notices that the first deep minimum was not significantly affected by the extrapolation, whereas the medium- and long-range behavior actually was to an appreciable extent. In fact, at variance with the $N_{conf} = 10^5$ estimate which flattens off over the largest sampled distances at a value of about −0.4 entropy units, the extrapolated function has already decayed to zero over distances beyond the third coordination shell. As for the resulting orientational pair entropy, we obtained—upon integrating $\tilde{I}_S^{(or)}(r)$—the value −11.5 e.u., an estimate that is fairly close to the one (−11.9 e.u.) that we inferred from the results reported in refs 15 and 16 at $T = 298$ K, which also were obtained without resorting to any approximate partial representation of the ODF. In any case, a

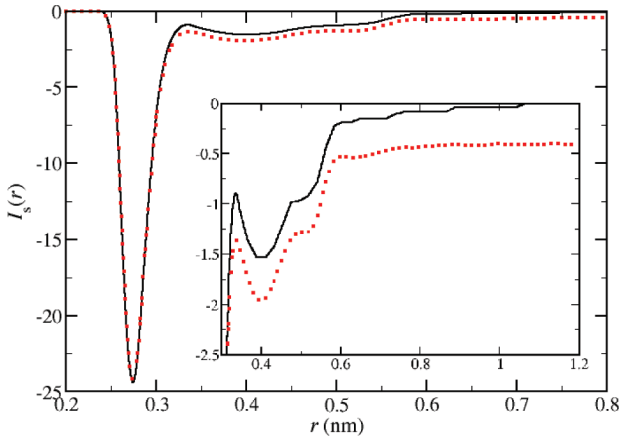


Figure 6. Integrand function, $I_S^{(gr)}(r)$, computed in the B_s set at ambient conditions: continuous (black) curve, asymptotic ($N_{conf} \rightarrow \infty$) estimate; dotted (red) curve, $N_{conf} = 10^5$ estimate.

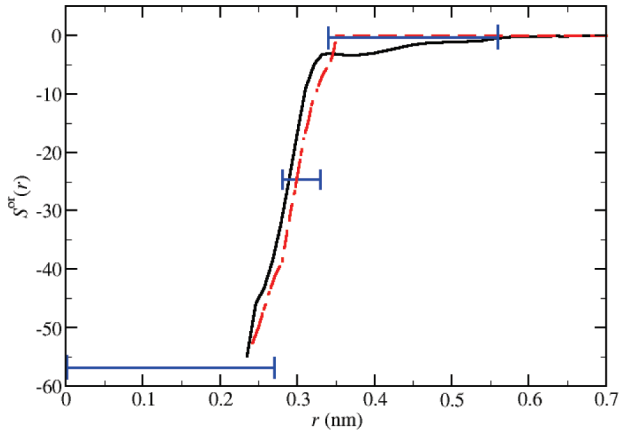


Figure 7. Orientational local entropy at ambient conditions: continuous (black) curve, present asymptotic estimate obtained in the B_s set; dot-dashed (red) curve, approximate three-shell AGP estimate (reproduced from Figure 4 of ref 9); horizontal (blue) bars, approximate three-shell GKSA estimates from ref 18.

modest increase of the orientational pair entropy such as the one we registered (∼0.4 e.u.) might be consistent with the 2 K temperature gap between the two calculations.

2. Approximate Results for the Orientational Pair Entropy. Figure 7 shows the outcome of our asymptotic extrapolation of $S^{(or)}(r)$ that is compared with the corresponding approximate estimate obtained by Lazaridis and Karplus using the AGP approximation.[9] We also included the OLE values (−56.77, −24.58, −0.34) that were obtained, in entropy units, from the orientational Shannon entropies of TIP4P water as calculated by Wang and co-workers (see Table 1 of ref 18) over three "representative" shells, i.e., $0 \leq r \leq 0.27$, $0.27 < r \leq 0.33$, and $0.33 < r \leq 0.56$, all distances being expressed in nanometers. We observe that both approximations miss the weakly modulated medium range tail of $S^{(or)}(r)$. Indeed, the modest global shift toward larger distances observed in the Lazaridis and Karplus estimate, as compared with the present calculation, clearly compensates for the faster and abrupt decay to zero of the function which, upon integration, does in fact lead to a result

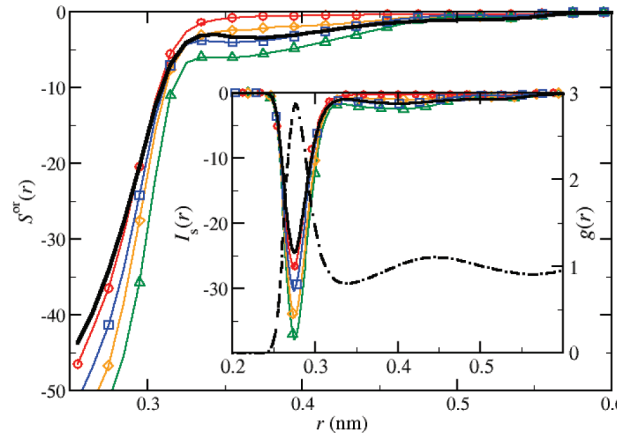


Figure 8. Orientational local entropy at ambient conditions: black curve, present asymptotic estimate obtained in the B_s set; red curve marked with circles, MSAGP estimate; blue curve marked with squares, MSAGP1 estimate; orange curve marked with diamonds, MSAGP2 estimate; green curve marked with triangles, MSAGP3 estimate. The inset shows the corresponding integrand functions $I_S^{(or)}(r)$ and the radial distribution function sampled over 10^5 configurations (right axis).

for the orientational pair entropy that is very close to the current estimate (see Table 1). A similar performance of the three-shell AGP approximation, with a partial yet fortuitous error compensation, was also observed in the other two thermodynamic states that we shall discuss in the following sections. On the other hand, the estimate reported by Wang and co-workers is one entropy unit larger than the present one, a discrepancy that is less obvious to explain since the method used by these latter authors was not based on a partial or full calculation of the ODF histogram, as done in ref 9 as well as in the present work.

The OLEs generated by the highest resolution ($\Delta R_{max} = \Delta r$) multishell variants of the AGP approximation introduced in section II.B are shown in Figure 8 for a reduced set of configurations ($N_{conf} = 10^4$). We first note that the multishell implementation of the original AGP approximation definitely improves over the original three-shell AGP approximation at short distances. In fact, this more spatially refined estimate closely reproduces the rise of the function for increasing r: the curve neither shows the rigid shift that we commented on above nor the kink at $r = 0.34$ nm, which is actually an artifact of using marginals averaged over discrete regions. Note, however, that the MSAGP approximation still fails to reproduce the tail of the function. Somewhat paradoxically, notwithstanding the improvement observed in the OLE at short distances, the multishell implementation of the AGP approximation leads to a lower estimate (in absolute value) of the orientational pair entropy (see Table 2) as compared with that obtained from the corresponding three-shell version (see Table 1). The worsened agreement is the consequence of the no longer fortuitously compensating failures which affected the short-range and long-range parts of $S^{(or)}(r)$ as estimated by Lazaridis and Karplus.

Better results on the long-range decay can be obtained by suitably modifying the AGP approximation so as to take into account missing angular correlations that plausibly play a

Table 2. Multishell AGP Estimates of the Orientational Pair Entropy (e.u.) of TIP4P Water Obtained, at Different Temperatures and Pressures, upon Sampling 1d and 2d Marginals over 10^4 Configurations, Using the Largest Number of Shells Compatible with the Spatial Resolution of the Calculation ($\Delta R_{max} = \Delta r = 0.01$ nm)[a]

approximation	I[b]	II[c]	III[d]
MSAGP	−11.7	−9.6	−9.0
MSAGP1	−15.8	−12.6	−12.7
MSAGP2	−16.6	−13.6	−13.0
MSAGP3	−21.2	−17.0	−17.4
Simulation	−14.8	−11.5	−11.3

[a] The asymptotic estimates generated by the extrapolated simulation data are also included for comparison. [b] $T = 260$ K, $P = 1$ bar ($R_{max} = 1.20$ nm). [c] $T = 300$ K, $P = 1$ bar ($R_{max} = 1.20$ nm). [d] $T = 300$ K, $P = 4$ kbar ($R_{max} = 1.15$ nm).

Table 3. Convergence of the Multishell AGP Approximations with Increasing Numbers of Configurations at $T = 300$ K and $P = 1$ bar

	number of configurations		
approximation	1×10^3	5×10^3	10×10^3
MSAGP	−9.74	−9.64	−9.62
MSAGP1	−12.91	−12.69	−12.63
MSAGP2	−13.82	−13.63	−13.60
MSAGP3	−17.44	−17.09	−17.02

role at larger distances. As discussed in section II.B, we tested three differently augmented AGP schemes. The results are shown in Figure 8. It clearly emerges that the MSAGP1 approximation is closer to the asymptotic MD estimate, from which it significantly deviates at very short distances only. The deeper minimum in $I_S^{(or)}(r)$ substantially accounts for the 10% discrepancy between the MSAGP1 and simulation estimates of $s_2^{(or)}$ (see Table 2). We recall that this particular scheme includes the *intramolecular* correlation between the angles θ and χ through the marginal distributions $g(\theta_1, \chi_1)$ and $g(\theta_2, \chi_2)$ whose role and importance have been recently highlighted by Wang and co-workers.[18] However, their approximation as well as the F7 factorization scheme also include cross-*intermolecular* correlations between the same angular pairs, a combination that we also exploited in the MSAGP3 approximation. As seen from Figure 8, the inclusion of the marginal distributions $g(\theta_1, \chi_2)$ and $g(\theta_2, \chi_1)$ worsens the agreement with the simulation data in that these additional correlations do actually overemphasize the structure of $S^{(or)}(r)$ both at small and large distances, producing an even more profound minimum in $I_S^{(or)}(r)$ as well as a more prominent long-range tail. Correspondingly, the estimated orientational pair entropy drops by more than four entropy units. The major responsibility of intermolecular (θ, χ) correlations in hollowing out a deeper minimum in the integrand function is also confirmed by the outcome of the MSAGP2 approximation where intramolecular (θ, χ) correlations have been neglected.

A significative property shared by all the multishell AGP approximations discussed above is their fast convergence rate as a function of the number of configurations. As seen from Table 3, the estimates of $s_2^{(or)}$ obtained after averaging the marginals over 5×10^3 configurations do in fact coincide with those produced with 10^4 configurations to the first decimal place.

B. TIP4P Water Close to the Temperature of Maximum Density. Upon lowering the temperature while keeping the pressure fixed at 1 bar, TIP4P water first exhibits the well-known maximum density anomaly at $T_{TMD} = 253 \pm 5$ K before congealing into ice I_h at $T_f = 232 \pm 5$ K, i.e., about 40 K below the experimental freezing point.[25] At 260 K, the average values of the specific density and excess internal energy were found to be 1.001 g cm^{-3} and − 10.67 kcal mol^{-1}, respectively. At lower temperatures, both the positional and angular order are more enhanced and longer-ranged. The RDF of the liquid is definitely more structured than at ambient conditions (see Figure 1), and we consistently found a value for the translational pair entropy (− 3.59 e.u.) that is about 18% lower than that obtained at 300 K. As for the calculation of the orientational pair entropy, we verified that the fit of $\sigma(r; N_{conf})$ was of comparable accuracy to that achieved at higher temperatures. Also in this case, the tail of $\sigma(r; N_{conf})$ turned out to scale as N_{conf}^{-1} at large distances. Figure 9 shows the extrapolated OLE and the corresponding integrand function. The comparison with the approximate estimates obtained from the four AGP schemes that we have already illustrated in the preceding sections confirms that even at this lower temperature the MSAGP1 approximation more faithfully reproduces the profile of $S^{(or)}(r)$, both at short and large distances. Correspondingly, the MSAGP1 estimate of the orientational pair entropy was again found to be closer than the other three approximate estimates to the asymptotic simulation value, the relative discrepancy being about 7% (see Table 2). We observe that $s_2^{(or)}$—and, correspondingly, the amount of angular order in water—is more significantly affected than $s_2^{(tr)}$ by the 40 K temperature drop. In fact, the value of the orientational pair entropy at 260 K was found to be about 29% lower than that at 300 K.

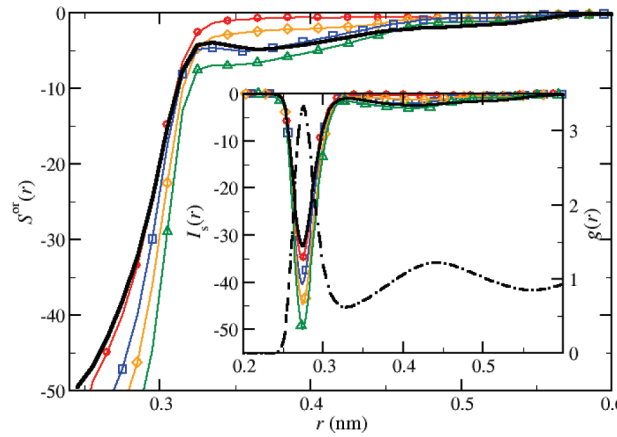


Figure 9. Orientational local entropy at $P = 1$ bar and $T = 260$ K: black curve, present asymptotic estimate obtained in the B_s set; red curve marked with circles, MSAGP estimate; blue curve marked with squares, MSAGP1 estimate; orange curve marked with diamonds, MSAGP2 estimate; green curve marked with triangles, MSAGP3 estimate. The inset shows the corresponding integrand functions $I_S^{(or)}(r)$ and the radial distribution function sampled over 10^5 configurations (right axis).

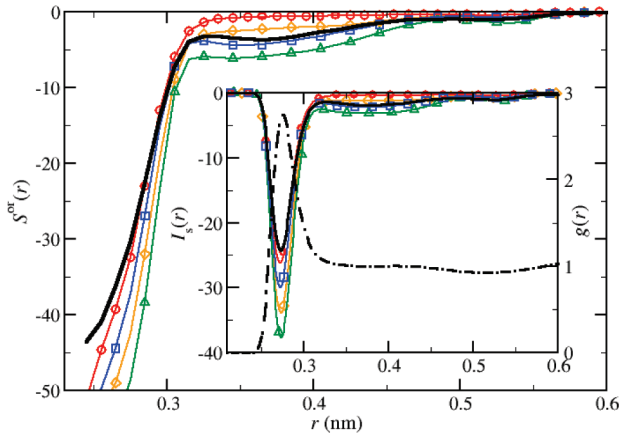


Figure 10. Orientational local entropy at $P = 4$ kbar and $T = 300$ K: black curve, present asymptotic estimate obtained in the B_s set; red curve marked with circles, MSAGP estimate; blue curve marked with squares, MSAGP1 estimate; orange curve marked with diamonds, MSAGP2 estimate; green curve marked with triangles, MSAGP3 estimate. The inset shows the corresponding integrand functions $I_S^{(or)}(r)$ and the radial distribution function sampled over 10^5 configurations (right axis).

C. TIP4P Water at Higher Pressure. We finally investigated the properties of ambient temperature TIP4P water compressed at a higher pressure ($P = 4$ kbar), falling in the range where crystalline ice II and ice III phases become stable at lower temperatures.[11,12] We found 1.134 g cm^{-3} and $-$ 10.18 kcal mol^{-1} for the average values of the specific density and excess internal energy, respectively. At variance with the behavior ordinarily observed in simple atomic fluids, the compression largely disrupts the local order observed in water at lower pressures. The effect on positional correlations is manifest in Figure 1. Notwithstanding this "antagonist" role played by the pressure, we found that, upon increasing P from 1 bar to 4 kbar at 300 K, the translational pair entropy dropped from -3.05 e.u. to -3.35 e.u.; the disruption of a relatively ordered network, which would imply a higher entropy, is more than compensated in this case by the reduction of available positional states produced by the 15% increase of the specific density. The decorrelating effect produced by compression is even stronger on angular order, as witnessed by the moderate increase that was registered instead in the orientational pair entropy (see Table 2), notwithstanding the increase of the density.[14]

The fit of $\sigma(r; N_{conf})$ turned out to be as accurate as that accomplished in the other two thermodynamic states. The scaling of the function as N_{conf}^{-1} at large distances was also confirmed. Figure 10 shows the extrapolated OLE and the corresponding integrand function. The comparison between the results obtained from the four AGP schemes confirms once more that even at higher pressures the MSAGP1 approximation is the most reliable approximation at short as well as large distances and also provides the most accurate estimate of the orientational pair entropy.

D. Comparison with Experimental Data. Table 4 presents a comparison between the cumulative pair entropies obtained from the current MD simulations without resorting to any approximation and the excess entropies of both TIP4P and ordinary water. The TIP4P values[26] of S_{ex} were obtained using thermodynamic integration methods for the calculation of the free energy, as extensively discussed in ref 12, while the experimental values follow from the data for the absolute entropy tabulated in ref 27, after subtracting the ideal-gas entropy. This latter contribution was calculated from eq (6.5) of ref 27, which parametrizes the ideal Helmholtz free energy. Note that the experimental estimate of the excess entropy at 260 K obviously refers to metastable undercooled water and was obtained upon extrapolating the values of the specific density and of the absolute entropy below the freezing temperature. We first observe that the pair entropy decreases upon lowering the temperature. The effect produced by an increase of the pressure on the local order of the liquid is more subtle, as already discussed above and more systematically analyzed in ref 14. In this specific instance, the effect is almost null, since the difference of 0.1 e.u. is presumably within the numerical uncertainty of the calculation. It should further be noted that the almost equal values found for S_2 at 300 K across the 4 kbar pressure gap is the outcome of differing relative weights of the translational and orientational pair entropies.

Table 4. Pair and Excess Entropies

		entropies (e.u.)		
T (K)	P (bar)	S_2 [TIP4P]	S_{ex} [TIP4P][a]	S_{ex} [expt.][b]
260	1×10^0	-18.39	-17.00	-15.73[c]
300	1×10^0	-14.57	-14.61	-13.99
300	4×10^3	-14.68	-15.14	-14.38

[a] Data from ref 26, reported with an estimated uncertainty of 0.06 e.u. [b] Estimates obtained from the data available in ref 27. [c] Estimate obtained upon extrapolating the properties of liquid water below the freezing point.

As shown in Table 4, at 260 K, the pair entropy actually overcomes the excess entropy: this implies a *positive* value of the so-called "residual multiparticle entropy" (RMPE), a quantity defined as the difference between S_{ex} and S_2.[19] As diffusely documented in the literature, a positive RMPE is evidence of a highly structured liquid.[13,28] On the other hand, at 300 K, the RMPE of TIP4P water almost vanishes at ambient pressure conditions, while being negative at 4 kbar but less than 3% of the total excess entropy. We remark that, as previously noted by Lazaridis and Karplus,[9] a small value of the RMPE does not necessarily imply that triplet or higher-order correlations do not play a role in determining the microscopic structure of the liquid. In fact, their overall contribution to the configurational entropy of a given substance may well be small or may even sum up to zero in some thermodynamic points or regions of the phase diagram, despite the fact that distribution functions beyond the pair one do not trivially reduce to the mere product of lower-order distribution functions.

It appears from Table 4 that at 300 K the pair entropy of TIP4P water provides fairly good estimates of the excess entropy of ordinary water. However, we are also aware that the agreement between the model and experimental data might be partially biased by the 40 K "shift" toward lower temperatures of the phase diagram predicted by the TIP4P model of water relatively to that of ordinary water.

V. Concluding Remarks

In this paper, we have presented a molecular dynamics calculation of the pair entropy of liquid water, modeled with the four-point transferable intermolecular potential (TIP4P) at three distinct thermodynamic states, corresponding to different values of temperature and pressure. The pair entropy is an integrated measure of two-body density correlations and represents the predominant contribution to the configurational entropy of a liquid. As such, it can be confidently used as a local structural estimate of the total configurational entropy: local, in that it does not call for an integration of the properties of the liquid along a thermodynamic path; structural, since it provides a direct connection between entropy and spatial order as monitored by the pair distribution function. This is the first aspect that we have put under scrutiny in this paper by extending preexisting analyses carried out for TIP4P water to thermodynamic states other than ambient conditions. Our new results for the orientational pair entropy follow from the calculation of the five-dimensional histogram that we obtained upon sampling, at given temperature and pressure, the configurations corresponding to different relative orientations of a generic water molecule with respect to a reference one, while keeping their centers of mass at a fixed distance. An intrinsic bias on the present results comes from the histogram bin width, with particular regard to the angular resolution. Our choice (10°), while being that commonly made in previous works on this subject, arises from a compromise between resolution and statistical quality of the calculations, a compromise that is unavoidably forced by the need for maintaining the overall size of the computation at a feasible level.

A secondary goal of this paper was that of discussing and testing some approximate schemes for the orientational distribution function that are based on the calculation of lower-order marginals, given the heavy computational task one has to face in a full-size calculation. In this respect, we have analyzed the performance of a number of factorizations that partially modify the "adjusted gas phase approximation" originally proposed by Lazaridis and Karplus.[9] We found that the best results, as far as both the orientational local entropy and its integrated value are concerned, were obtained at all of the three sampled thermodynamic states when using the so-called MSAGP1 approximation, which includes intramolecular correlations between the angle formed by the dipole vector of a water molecule and the intermolecular axis, and the angle describing the rotation of the same molecule about its own dipole vector. We emphasize that the MSAGP1 estimates were obtained upon sampling the marginal distribution functions over 10^4 configurations only and appear to underestimate the corresponding "exact" molecular dynamics values—which were obtained with a sampling carried out over a 10 times larger number of configurations—by 12% at the most.

On the other hand, we have also verified that including the intermolecular contribution which arises from cross correlations between the same angles mentioned above but measured on different molecules does actually worsen the agreement with the molecular dynamics results in that the ensuing modified scheme (MSAGP3) manifestly overestimates the degree of orientational order present in the liquid in a systematic way.

Acknowledgment. E.G. acknowledges a Ph.D. grant from CNISM (Italy) and the scientific hospitality of the Université Pierre et Marie Curie (Paris, France) in 2008 during a six-month visit cofunded by the Erasmus/Socrates program and also thanks Drs. Rubens Esposito and Dino Costa for numerical assistance at different stages of this work. P.V.G. thanks Prof. Carlos Vega for providing some reference thermodynamic data on the TIP4P model and Dr. Lingle Wang for some clarifying comments on the calculation of the orientational entropies with the *k*th nearest-neighbor method. The authors gratefully acknowledge the allocation of computer time from the French National Supercomputing Facility IDRIS, within the projects CP9-81387 and CP9-91387, and from the "Centro di Calcolo Elettronico "Attilio Villari" della Università degli Studi di Messina".

References

(1) Ben-Naim, A. *Molecular Theory of Water and Aqueous Solutions - Part 1: Understanding Water*; World Scientific: Singapore, 2009.

(2) Ball, P. *Nature* **2008**, *452*, 291.

(3) Green, H. S. *The Molecular Theory of Fluids*; North-Holland: Amsterdam, The Netherlands, 1952.

(4) Stratonovich, R. L. *Sov. Phys. JETP* **1955**, *28*, 409.

(5) Nettleton, R. E.; Green, M. S. *J. Chem. Phys.* **1958**, *29*, 1365.

(6) Yvon, J. *Correlations and Entropy in Classical Statistical Mechanics*; Pergamon Press: Oxford, United Kingdom, 1969.

(7) Baranyai, A.; Evans, D. J. *Phys. Rev. A* **1989**, *40*, 3817.

(8) Prestipino, S.; Giaquinta, P. V. *J. Stat. Phys.* **1999**, *96*, 135; **2000**, *98*, 507.

(9) Lazaridis, T.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 4294.

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

(11) Sanz, E.; Vega, C.; Abascal, J. L. F.; MacDowell, L. G. *J. Chem. Phys.* **2004**, *121*, 1165.

(12) Vega, C.; Sanz, E.; Abascal, J. L. F.; Noya, E. G. *J. Phys.: Condens. Matter* **2008**, *2*, 153101.

(13) Saija, F.; Saitta, A. M.; Giaquinta, P. V. *J. Chem. Phys.* **2003**, *119*, 3587.

(14) Esposito, R.; Saija, F.; Saitta, A. M.; Giaquinta, P. V. *Phys. Rev. E* **2006**, *73*, 040502(R).

(15) Zielkiewicz, J. *J. Chem. Phys.* **2005**, *123*, 104501.

(16) Zielkiewicz, J. *J. Chem. Phys.* **2006**, *124*, 109901.

(17) Zielkiewicz, J. *J. Phys. Chem. B* **2008**, *112*, 7810.

(18) Wang, L.; Abel, R.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2009**, *5*, 1462.

(19) Giaquinta, P. V.; Giunta, G. *Physica A* **1992**, *187*, 145.

(20) Giaquinta, P. V.; Giunta, G.; Prestipino Giarritta, S. *Phys. Rev. A* **1992**, *45*, R6966.

(21) Jakse, N.; Charpentier, I. *Phys. Rev. E* **2003**, *67*, 061203.

(22) Prestipino, S.; Giaquinta, P. V. *J. Stat. Mech.* **2004**, P09008.

(23) Martyna, G. J.; Tuckerman, M. E.; Klein, M. L. PINY_MD (c) Simulation Package, 2002.

(24) *IUPAC Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book"). XML on-line corrected version: http://goldbook.iupac.org/E02151.html (accessed Nov 23,2009).

(25) Vega, C.; Abascal, J. L. F. *J. Chem. Phys.* **2005**, *123*, 144504.

(26) Vega, C. 2009, private communication.

(27) Wagner, W.; Pruss, A. *J. Phys. Chem. Ref. Data* **2002**, *31*, 387.

(28) Giaquinta, P. V. Entropy revisited: The interplay between entropy and correlations. In *Highlights in the Quantum Theory of Condensed Matter*; Beltram, F., Ed.; Publications of the Scuola Normale Superiore (Selections), Edizioni della Normale: Pisa, Italy, 2005; Vol. 1, pp 9−14. See also the introduction and the bibliography in Giaquinta, P. V. *Entropy* **2008**, *10*, 248.

CT900627Q

JCTC Journal of Chemical Theory and Computation

Stress Analysis at the Molecular Level: A Forced Cucurbituril-Guest Dissociation Pathway

Michael K. Gilson*

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850

Received December 14, 2009

Abstract: Changes in mechanical stresses in a tight-binding host−guest system were computed and visualized as the cation was computationally pulled out of the cucurbituril host in a series of steps. A sharp conformational transition was observed as one of the guest's ammonium groups jumped through the center of the host to the opposite portal. The conformation immediately prior to this transition was found to possess high levels of Lennard-Jones and electrostatic stress. This observation, along with the specific distribution of Lennard-Jones stress around the portals, suggested that the conformational transition resulted from steric constriction, which had been expected, and electrostatics, which was not expected. An important role for electrostatics, at least at the level of these calculations, was confirmed by a comparative computational pulling study of another guest molecule lacking the critical ammonium group. These calculations suggest that the binding kinetics of diammonium guests that position an ammonium at each cucurbituril portal will be found to be slower than the kinetics of monoammonium guests. More generally, the results suggest that computational stress analysis can provide mechanistic insight into supramolecular systems. It will be of considerable interest to extend such applications to biomolecules, for which the mechanisms of conformational change are of great scientific and practical interest.

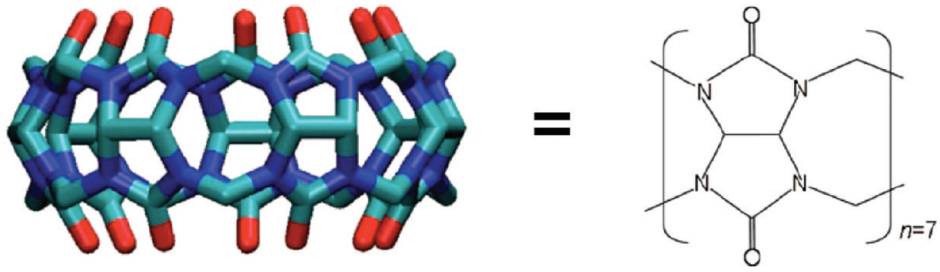
Introduction

The concepts of mechanical stress and strain are widely used. For example, engineers rely on them when designing structures, and geophysicists study stress waves propagating through the earth. Materials scientists have, furthermore, developed computational methods of extracting measures of stress from atomistic simulations,[1−6] and such approaches have recently begun to find application at the level of single molecules. Thus, Rafii-Tabar has discussed stress and strain in carbon nanotubes,[7] and Yamato and co-workers used simulations to observe the propagation of stress in photoactive yellow protein during the "protein quake" generated by photoisomerization of the chromophore.[8] In related earlier work, Yamato and co-workers computed molecular strain in proteins under hydrostatic pressure,[9] and Beuhler and coworkers have recently computed stress in connective tissue proteins.[10] Such applications suggest that atomistic stress theory can contribute to our understanding of molecular mechanisms and perhaps even guide molecular design. However, stress calculations have not, so far, been applied in the context of supramolecular chemistry.

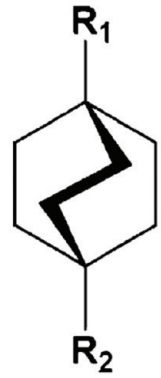
The cucurbiturils,[11,12] chemical hosts constructed as rings of glycouril monomers (Figure 1), have been the subject of increasing interest in recent years, as they are relatively easy to synthesize and display distinctive molecular recognition properties, including the ability to bind dicationic guests from aqueous solutions,[13] sometimes with affinities rivaling the tightest protein−ligand systems.[14−16] The cucurbiturils have a rich range of properties and potential applications, as recently reviewed.[17−19] Examples include the formation of polyrotaxanes,[20,21] modulation of the fluorescence of guest dye molecules,[22] chemical catalysis,[23] and, for specialized variants, two-site allosteric binding.[24] The mechanisms and

* Corresponding author. Current address: Skaggs School of Pharmacy and Pharmaceutical Sciences, 9500 Gilman Drive MC 0736, La Jolla, CA 92093-0736. Phone: 858-822-0622. E-mail: mgilson@ucsd.edu.

10.1021/ct900668k
Published on Web 02/22/2010



Cucurbit[7]uril



B4: R_1=$CH_2NH_3^+$, R_2 = H.
B5: R_1=R_2= $CH_2NH_3^+$

Guests B4, B5.

Figure 1. Host cucurbit[7]uril host (CB7) and guests B4 (monoammonium) and B5 (diammonium).

kinetics of cucurbituril−guest binding are thus of considerable interest. It is already known that the size of the guest molecule can strongly influence binding kinetics for cucurbit[6]uril (CB6) in a manner that is essentially independent of binding thermodynamics.[25,26] This is presumably because its portals are of smaller diameter than its internal cavity. Such constrictive binding[27,28] highlights the importance of steric interactions as determinants of binding kinetics for these systems, but the pH dependence of binding kinetics for titratable guests with CB6 indicates that electrostatic interactions also can influence the stability of the transition state.[29,26]

Here, we use computer modeling to study the causes and distributions of atomistic stress in a host and guest when they are forced apart, as if in an atomic force microscopy (AFM) single-molecule experiment. The system consists of cucurbit[7]uril (CB7)[11,12] with a dicationic compound (B5) predicted[30] and proven (Inoue et al., to be published) to bind CB7 with an ultrahigh affinity similar to that previously observed for a dicationic ferrocene derivative.[15] We also carry out comparative calculations for a monocationic guest (B4) predicted[30] to bind CB7 less tightly by 3 kcal/mol than B5 (see Figure 1.) Novel methods of computing and visualizing changes in atomistic stress are used to evaluate the relative roles of steric constriction and electrostatics as kinetic determinants in these systems and reveal informative and unexpected distributions of molecular stress as the complexes dissociate.

Methods and Concepts

Molecular Modeling of a Forced Host−Guest Dissociation. The calculations started with the most stable conformation of the host−guest complex identified with the M2 free energy method,[31,32] based on previously described force-field and implicit solvation parameters.[30] We judged a nitrogen of the guest and an equatorial carbon of the host to be chemically plausible points for added linkages to an AFM tip and a fixed substrate and assigned pulling forces to these atoms in a computationally simple fashion by adding an artificial harmonic bond between them, as illustrated in Figure 2. The N−C distance was about 5.5 Å in the initial conformation of the complex. The equilibrium length B_o of the artificial bond, with force constant 100 kcal/mol/Å^2, was increased in increments, and gradient-based energy minimization was carried out so that the artificial bond would drive the guest molecule partly or fully through and out of the host cavity. The length of the artificial bond after minimization, B, was recorded and was also used to compute the magnitude of the force, F, exerted by the artificial bond at its N and C attachment points, where $F = 100(B - B_0)$. This procedure was repeated for various equilibrium lengths of the artificial bond, and stresses were calculated for the

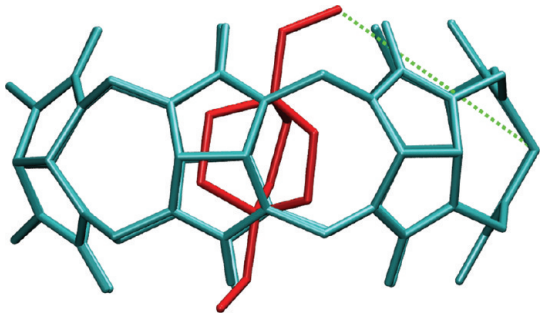


Figure 2. Stable bound conformation of CB7 (cyan) with B5 (red), with an artificial forcing bond (green dotted line). Hydrogen atoms omitted.

resulting conformations. These stresses were compared with the stresses computed for the most stable conformations of the free host and guest with the M2 method. The stress calculations are described below. Hydration effects were treated approximately by a generalized Born model.[33−35] Because the present calculations do not account for thermal motion, barrier-crossing results only from the conformational forcing of the artificial bond, rather than from random thermal motion, and the calculations model only one dissociation pathway. They thus model a fast pulling process, and the forces and stresses are expected to be larger than those seen at the lower pulling speeds of typical AFM experiments. A similar forcing process has been used in a prior computational examination of cucurbituril−guest dissociation.[26]

Mechanical Stress and Its Calculation and Visualization at the Atomistic Level. The stress as a function of position within an object is a tensor field that describes the action of local forces. Tensile stress occurs when a volume element of material is pulled in opposite directions from opposite sides, compressive stress when it is pressed on from opposite sides, and shear stress when opposite sides are subjected to oppositely directed tangential forces. A volume element can be under stress yet experience zero net force. Accordingly, an object can be in mechanical equilibrium, yet internally stressed. For example, the cables of a suspension bridge are under tensile stress, and its towers are under compressive stress. Furthermore, a local force on a body can place the whole under stress, with long-ranged consequences. The transmission of torsional stress and its consequences for global DNA supercoiling provides an example at the molecular level. The stress on a volume element generates a deformation, i.e., to strain, which is also a tensor field. For elastic materials, stress and strain are linearly related to each other, and the energy of elastic deformation is directly related to the product of stress and strain. Such a linear relationship is not expected to hold in general for molecules, although it may be a good approximation for ones that are relatively stiff. Note that strain in this mechanical sense is not identical with "chemical strain": mechanical strain is defined as a spatial deformation induced by stress,[36] as just described, whereas chemical strain is the enthalpy of a molecule relative to an unstrained reference structure.[37]

Here, we computed atomistic stresses for individual energy-minimized conformers of the host, the guests, and their complexes. Following Zimmerman and co-workers,[6] we use Hardy's expression for the local stress tensor.[38] The stress tensor σ_i at each atom i was thus computed as

$$\sigma_i = \frac{1}{v} \sum_{\text{local}} r_{ij} \otimes f_{ij}$$

where j indexes atoms within a spherical region of volume v_{loc} local to atom i, r_{ij} is the vector from atom i to atom j, and f_{ij} is the force exerted by atom j on atom i. This formula is most straightforwardly applied to center-to-center forces, so attention here is limited to bonded, Lennard-Jones, Coulombic, and Generalized Born (GB) forces, the latter excluding the force contributions from variation in the effective Born radii. We found it informative to compute and visualize the stresses associated with each separate force term, combining the Coulombic and GB stresses to determine net electrostatic stress. The radius of the spherical region was set to 5 Å, and all noncovalent force calculations were cut off at this range. Covalent bond stresses accounted for only atoms j bonded directly to atom i. Here negative and positive values of σ_i imply tensile and compressive stresses, respectively. For constrictive host−guest binding, one may anticipate that forcing a bulky guest out through a narrow portal will generate compressive Lennard-Jones stresses between the guest and the portal, along with tensile bond stresses around the portal itself.

Initial calculations indicated that the free molecules possessed significant baseline stress. We were interested in stress changes on binding and therefore wished to subtract the stresses of the free molecules from the stresses of their complexes. However, a naïve initial approach to evaluating these differences—simply subtracting the stress tensors—was unhelpful because changes in the lab-frame orientation of the molecules caused matching tensors in the bound and free states not to cancel. We therefore subtract the stress tensors in local molecular coordinate systems, as follows. We transform each atom's lab-frame principal stresses before and after binding into local Cartesian coordinates, subtract free from bound stresses in these local coordinates, then convert back to lab coordinates for analysis and display. Separate local Cartesian coordinates were set up for each atom i by choosing two atoms j and k, such that i is bonded to j and j to k. Then the local x axis was aligned with the ij vector, the local y axis was placed in the ijk plane and orthogonal to the x axis, and the z axis was oriented along the cross-product of the x and y unit vectors. The same atom triplets were used for the free and bound states, in order to establish consistent local coordinates.

Stresses, or stress differences, at each atom i were visualized by diagonalizing the tensor σ_i to provide the magnitudes and directions of its 3 principal stresses. The program VMD[39] was then used to render each principal stress as a spindle-shaped glyph made of two thin cones based at atom i, extending in the positive and negative directions along the direction of the principal stress and having a length proportional to the magnitude of the principal stress. Tensile (negative) and compressive (positive) stress components were distinguished by coloring their corresponding spindles green and orange, respectively. The length of each cone (Å) was

set to the magnitude of the corresponding principal stress (kcal/mol/Å), scaled by the factor $0.2v_{loc}$. This visualization approach differs from previously reported tensor glyphs designed for visualization of stresses in continuous media.[40−43]

Results

Stresses in the Unbound Host, CB7, and Guest, B5. Figure 3 displays the stresses computed for the most stable conformation found for the free guest B5. Bond-stretch stresses (top left) are tensile (orange spindles) in the bicyclooctane moiety and weakly compressive in the ammonium groups. The Lennard-Jones forces (top right) show compression (green spindles) in the bicyclooctane moiety, reflecting steric repulsion within this compact ring system and accounting for the tensile bond stresses. The Coulombic stresses (middle left) reflect chiefly the influence of the two ammonium groups, which repel the aliphatic hydrogens of the bicyclooctane moiety because of their weakly positive charge and therefore place them under compressive electrostatic stress. The weakly negative bicyclooctanes are correspondingly placed under tensile electrostatic stress. The GB stresses (middle right) largely cancel the Coulombic stresses when summed (bottom left), leaving mainly tensile stress on the ammonium groups, presumably due to unbalanced GB forces drawing them toward the high dielectric solvent. The summed bond-stretch, Lennard-Jones, Coulombic, and GB stresses (bottom right) do not fully cancel, indicating that this molecule is stressed in even this energy-minimized conformation.

Figure 4A and B display, respectively, top and side views of the stresses computed for the most stable conformation found for the free host, CB7. The bond stresses (top left), which are mainly tensile (orange), appear to result largely from local forces intrinsic to the system of linked glycuril rings. The Lennard-Jones forces (top right), which are mainly compressive and circumferentially oriented, chiefly reflect side-to-side repulsions between neighboring carbonyl moieties. The Coulombic (middle left) and GB stresses (middle right) appear more complicated and display larger radial components than the bonded and Lennard-Jones stresses, but as for the free guest (above), they largely cancel when summed (lower left). The sum of all computed stresses (lower right) shows considerable overall stress, including compression of the carbonyl carbons and tension of most other atoms in the glycouril units.

Stress in the Relaxed Starting Conformation of the Complex. The starting conformation of the CB7−B5 host−guest complex was generated by a conformational search and energy minimization in the absence of the artificial forcing bond. The guest's bicyclooctane moiety lies in the middle of the CB7 cavity, while its two ammonium groups sit at opposite portals of the host and donate hydrogen bonds to its carbonyl oxygens. The three panels in the left-hand column of Figure 5 visualize differences in stress, relative to the free host and guest, for this unforced, energy-minimized complex. The bond stresses (top left) have become less tensile and more compressive, relative to the free molecules, as indicated by the ring of green spindles around each of the portals of the host, with some variation

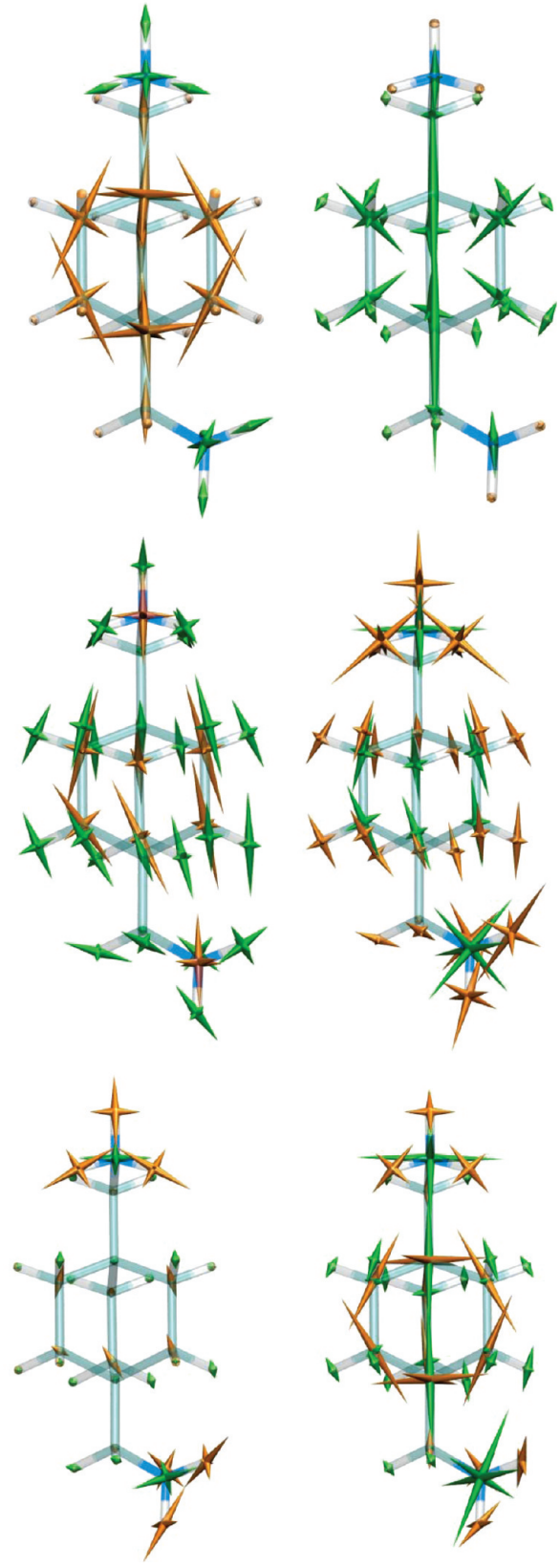


Figure 3. Stresses in free guest B5. Bonded: top left. Lennard-Jones: top right. Coulombic: middle left. Generalized Born: middle right. Electrostatic (Coulombic + GB): lower left. Total: lower right.

of stress around the ring. These observations are consistent with the fact that binding has reduced the diameter of the portals slightly and made the host less round and more oval

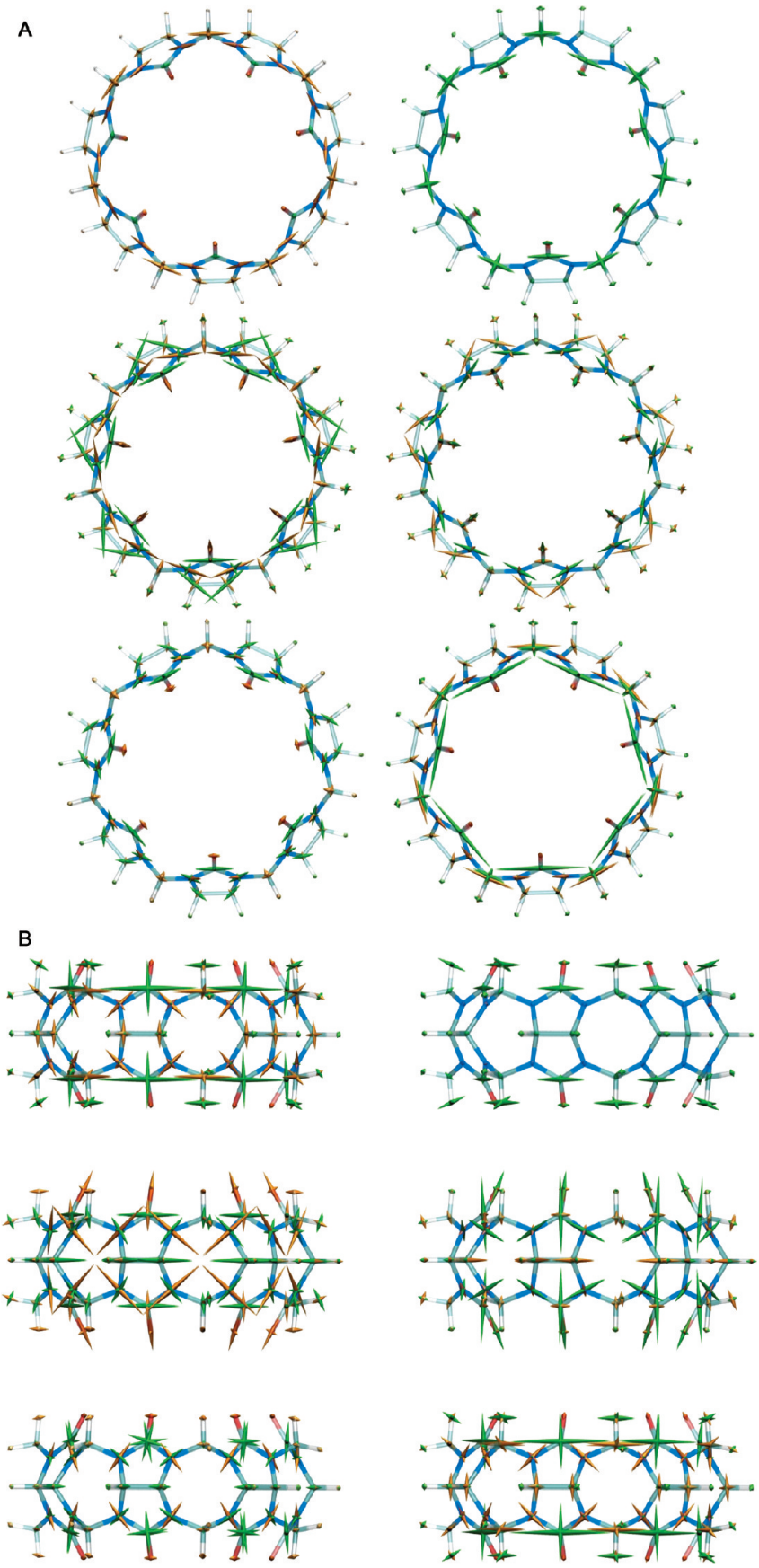


Figure 4. (A) Stresses in free host, CB7, top view. Bonded: top left. Lennard-Jones: top right. Coulombic: middle left. Generalized Born: middle right. Electrostatic (Coulombic + GB): lower left. Total: lower right. (B) Same as in A in side view.

in shape. One can also see an increase in the tensile bond stress on the ammonium groups (orange spindles), presumably due to their hydrogen bonding to the host's carbonyl oxygens. The Lennard-Jones stresses (middle left) change

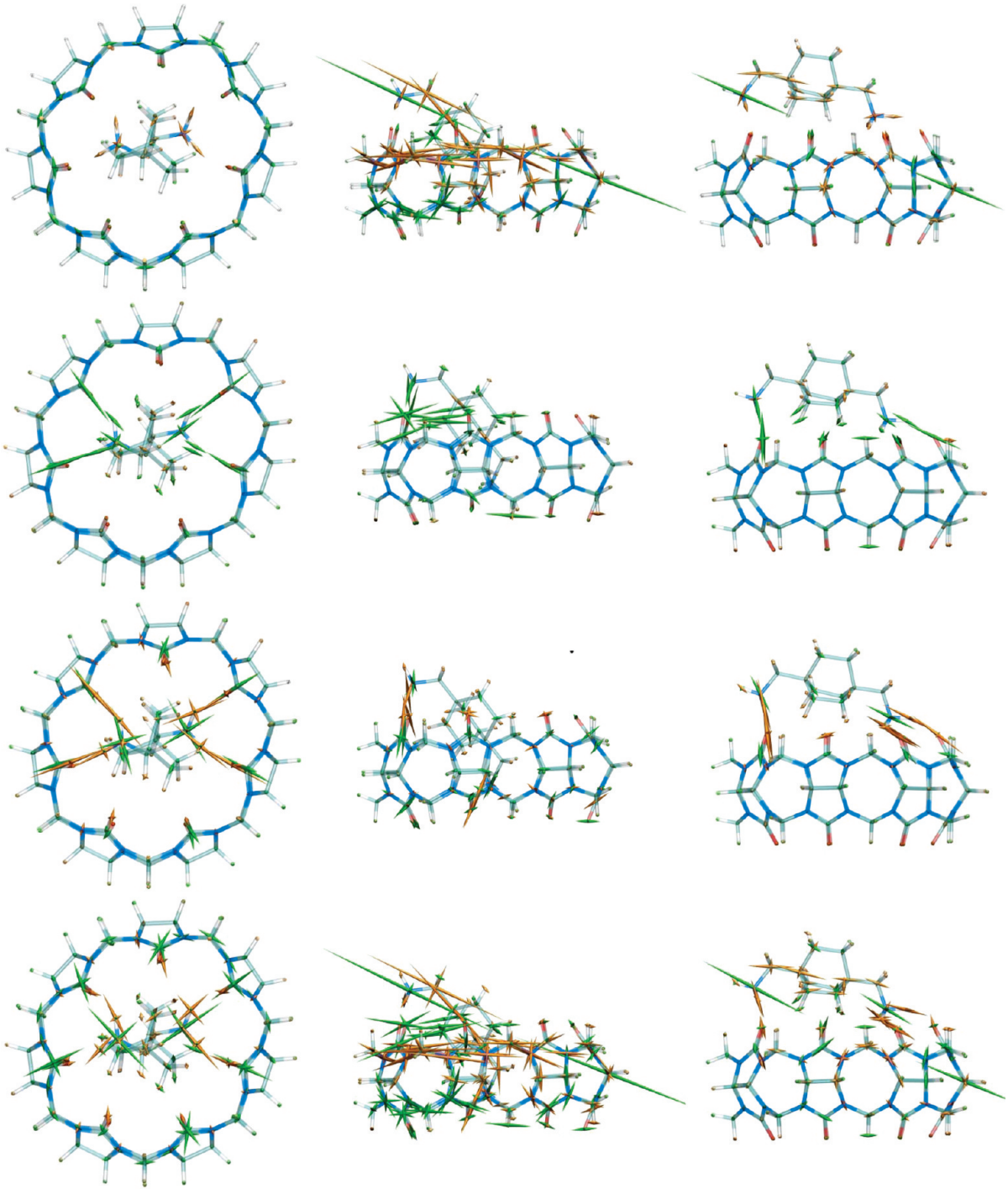


Figure 5. Stresses in three different conformations of the CB7 complex with diammonium guest B5. Left column: unforced starting structure ($B = 5.5$ Å), top view. Middle column: conformation with largest force imposed by the artificial bond ($B = 12.3$ Å), side view. Right column: conformation immediately after the conformational transition ($B = 12.4$ Å), side view. Bond stresses: top row. Lennard-Jones stresses: center row. Electrostatic (Coulombic + GB) stresses: bottom row. The lengths of the bond stress symbols on the two atoms involved in the artificial bond (Figure 2) were scaled down by an additional factor of 0.2.

little relative to the free molecules except for striking new compressive stresses involving the ammonium groups and the carbonyl oxygens to which they are hydrogen-bonded in the complex. These result from steric compression of the atoms involved in the hydrogen bonds. The same hydrogen bonds are also associated with strong electrostatic stresses (lower left): the attractive interactions of oppositely charged atoms involved in the hydrogen bonds (e.g., ammonium hydrogen and carbonyl oxygen) cause tensile stresses (orange), while the associated repulsions (e.g., ammonium nitrogen and carbonyl oxygen) lead to compressive electrostatic stresses (green).

A Sharp Conformational Transition. The guest is forced out of the host by incremental increases in the equilibrium length of the artificial bond, B_0 (Figure 2). The bottom ammonium moves to the top portal, while the top ammonium group remains on top. The resistance of the guest to exiting

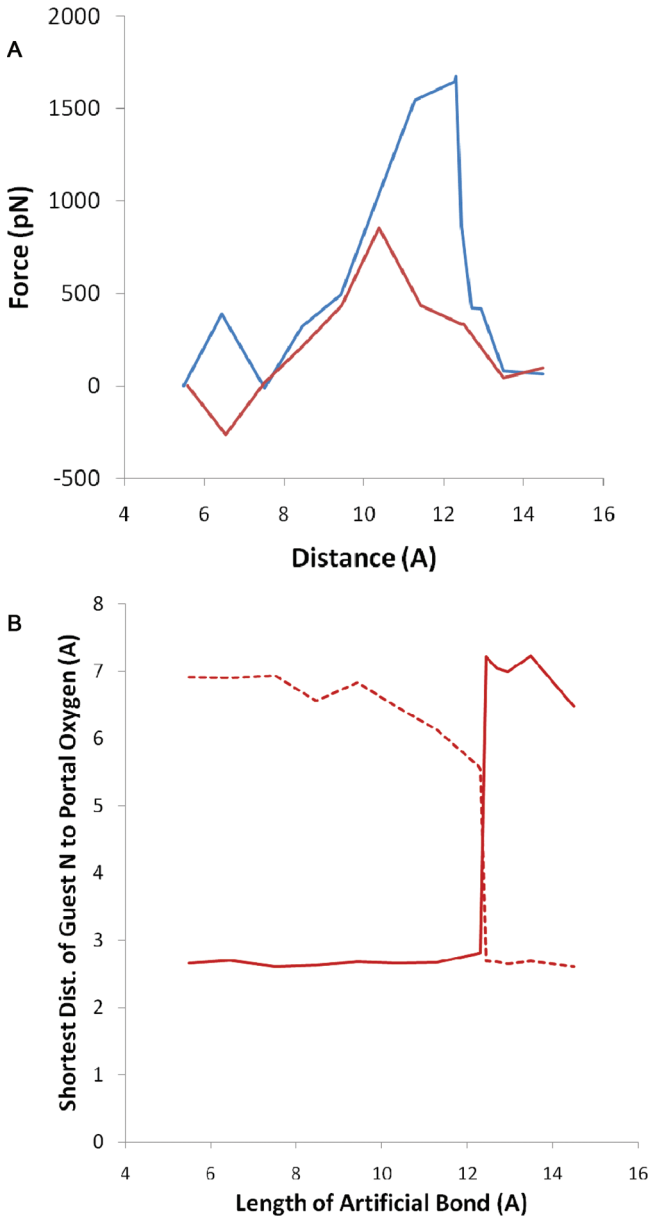


Figure 6. (A) Computed forces exerted by the artificial bond vs the bond length, for guests B5 (blue) and B4 (red). (B) Closest distance of bottom ammonium nitrogen to bottom portal (solid) and top portal (dashed), as a function of the length, *B*, of the forcing bond.

the host is manifested by the magnitude of the force, *F*, exerted by the artificial bond as a function of its length, *B*. As graphed in Figure 6A (blue), this force peaks at $B = 12.3$ Å and then drops sharply at 12.4 Å. The drop corresponds to a conformational transition in which the guest's bottom ammonium cationic group jumps from the bottom portal of the host to the top one. This jump is illustrated in Figure 6B, which shows a sudden increase in the distance of the bottom ammonium group from the bottom portal and a simultaneous drop in its distance to the top portal. The conformations immediately before and after this transition are shown in the middle and right-hand columns of Figure 5, respectively.

When the pulling force is at its maximum ($B = 12.3$ Å), the complex is highly stressed (Figure 5, middle column). The graphical visualization of bonded stress (top middle) includes two large, diagonal green spindles (rendered at reduced scale) at the N and C atoms used as attachment points for the artificial bond. These indicate that the artificial bond is compressed, because it is pushing the N and C atoms apart. There is also a large buildup of tensile bonded stress (top middle) around the top portal. The stress is greatest at the left of the top portal, toward which the guest is being pushed. In addition, and rather unexpectedly, the left side of the lower portal is under compressive, rather than tensile, stress (green). There is also tensile bonded stress along the long axis of the guest, consistent with the fact that its top ammonium is being pushed up and to the left while the rest of the guest is stuck partway through the host. The Lennard-Jones stresses (center middle) are also focused at the left-hand side of the top portal, since the artificial bond is pushing the guest against the host in this region. The electrostatic stresses (bottom middle) reflect hydrogen-bonding of the top ammonium with the top portal and also show tension between the bottom ammonium and the bottom portal.

After the conformational transition ($B = 12.4$ Å), when the bottom ammonium has jumped through the host and is now hydrogen-bonded with the top portal (Figure 5, right-hand column), the stresses have become markedly smaller. This is particularly evident in the bonded stresses (top right vs top middle). It is interesting to note a persistent spine of modest tensile stress along the axis of the guest molecule, presumably due to the interplay of the imposed pulling force and the restraining H bonds of the ammonium groups with the carbonyls. The Lennard-Jones stresses also have fallen markedly (center right) and show mainly compressive steric stresses, once again involving ammonium−carbonyl hydrogen bonds. The electrostatic stresses (bottom right) also are consistent with this pattern of hydrogen-bonding.

Physical Basis for the Conformational Transition. We had initially conjectured that the sharp conformational transition resulted from a build-up and release of compressive Lennard-Jones stress as the bulky bicyclooctane group passed through a constrictive exit portal, much as previously argued for a host−guest system involving cucurbit[6]uril (CB6).[26] However, this view was not fully supported by the stress analysis, because the compressive Lennard-Jones stress prior to the transition (Figure 5, top middle) does not extend around the whole portal but is localized on the left-hand side, where the artificial bond is driving the guest. In addition, the bicyclooctane group appears to have already passed most of the way through the top portal by time the forcing bond has reached a length of 12.3 Å. An alternative view is suggested by the observation of tensile electrostatic stresses between the bottom ammonium and the bottom portal (Figure 5, bottom middle). Both Coulombic attractions between the bottom ammonium and the bottom portal carbonyls and GB forces resisting desolvation of the ammonium group could contribute to the resistance of the bottom ammonium to passing through the host, and the view that electrostatics make a major contribution to the dissociation barrier would be broadly congruent with prior observations regarding the role of the guest molecule's pH-dependent charge in the binding kinetics of CB6.[29]

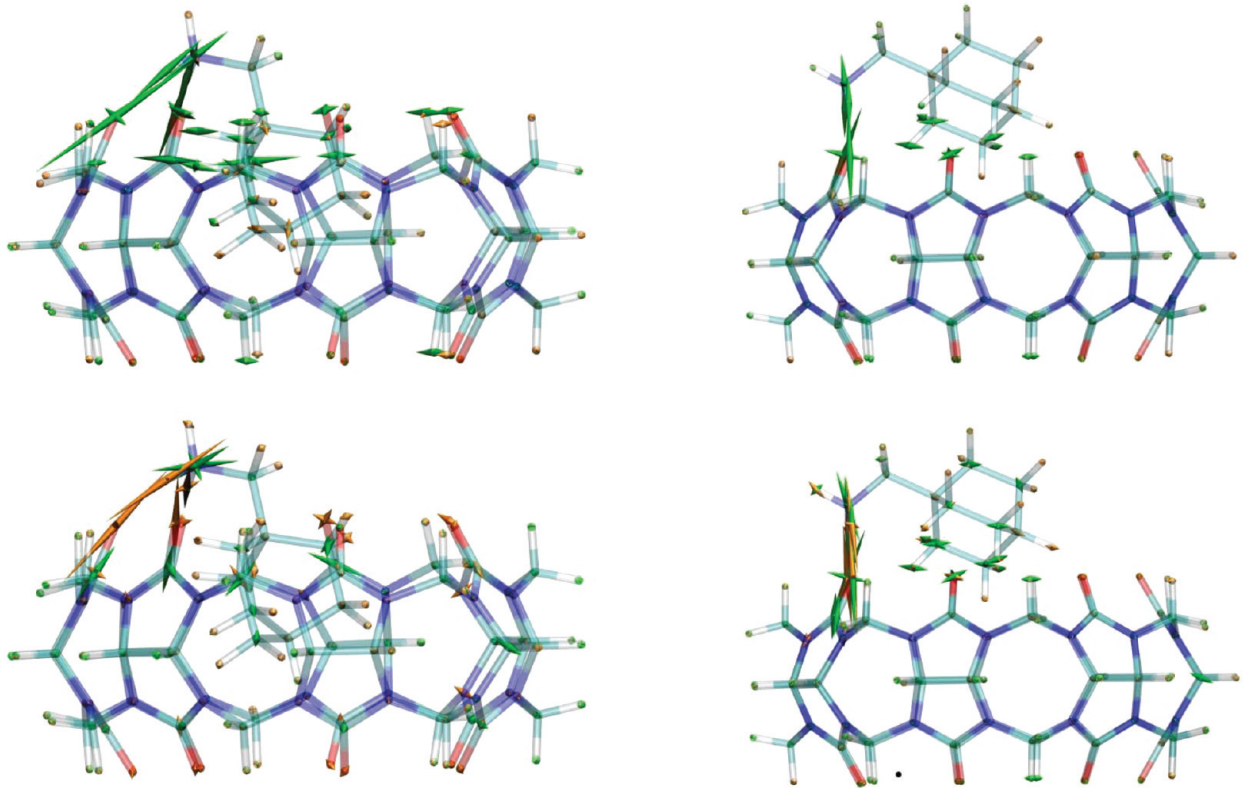


Figure 7. Lennard-Jones (top) and electrostatic (Coulombic + GB, bottom) stresses in two conformations of the CB7 complex with monoammonium guest B4. Left: conformation with peak force. Right: conformation immediately after peak force, with guest shifted to top portal.

We tested this idea by repeating the calculations with guest B4 (Figure 1), which is the same as B5 except that it lacks one methylammonium group. The computational pulling experiment was arranged so that the single ammonium group would lie at the top portal and only the bulky but electrically neutral bicyclooctane moiety would need to clear the exit portal. The peak of the new length-force curve (Figure 6a, red) is only half as high as the original and is shifted from 12.3 Å to 10.4 Å. This change indicates that the original electrostatic impediment to dissociation has been removed and its place taken by a different interaction. Visualization of the B4 complexes reveals that this guest, too, undergoes a sharp conformational transition in which the bicyclooctane moiety jumps from an intermediate location (Figure 7, left column) to the top portal (Figure 7, right column) as the peak in the length−force curve is passed. The pattern of Lennard-Jones stresses (Figure 7, top row) is similar to that for B5 at the peak of its force curve (Figure 5 top), but less intense. The same is true for the bond stresses (not shown). The electrostatic stresses (Figure 7, bottom row) are very similar to those for B5 (Figure 5, bottom), except for the absence of stresses on the deleted ammonium group. Comparison of these results with those for guest B5 support the view that electrostatics plays a major role in the conformational transition of the latter.

Discussion

This computational stress analysis yields mechanistic insights into the forced dissociation of a high affinity cucurbituril−guest complex. We had initially expected to observe a classic constrictive binding process, i.e., one in which a peak in the length−force curve resulted from steric hindrance as the bulky bicyclooctane moiety passed through a relatively narrow CB7 portal.[27,28] However, the pattern of Lennard-Jones stress at the peak of the length−force curve did not appear fully consistent with this expectation, and we observed, furthermore, that the ammonium being pulled through the host was under tensile electrostatic stress. These observations suggested that electrostatics might play an important role in establishing the peak in the force curve. Such an explanation would be physically reasonable because the hydrogen bond of the bottom ammonium to the bottom portal must be broken for the forced dissociation to occur. Moreover, the work of dehydrating the ionized ammonium on entering the host's cavity should be substantial. Comparative calculations for another guest molecule lacking the bottom ammonium group supported the importance of electrostatics, because the force maximum was reduced and shifted. These results suggest that the association kinetics of CB[7] guests will tend to be faster for monocationic guests than for diammonium guests that position on ammonium at each portal. Both binding and dissociation are expected to be slowed by the requirement of driving one of the ammonium groups through the middle of the host for either process. It is reasonable that electrostatics should be of central importance for diammonium guests, given that even monocationic guests have kinetics that are sensitive to charge.[29]

It was also interesting to observe a complex distribution of bond stresses around the portals of the host in the highly stressed conformation immediately before the conformational

transition involving the diammonium guest, B5. The tensile forces at the top portal could have been anticipated, but the appearance of compression around the bottom portal was more surprising. Such results are reminiscent of stress patterns computed for engineered structures and could be useful as a basis for molecular engineering aimed at tailoring energy barriers or molecular processes and pathways.

The present treatment of the forced dissociation process is preliminary in the sense that, like a prior study,[26] it does not account for thermal motion. The calculations thus are inconsistent with the quasi-adiabatic assumption that the system equilibrates rapidly with respect to the time constant for barrier crossing.[44] Therefore, we have in effect studied only a single dissociation pathway for the CB7−B5 complex. One way to arrive at a more comprehensive description might be to carry out a molecular dynamics simulation for each equilibrium length of the artificial forcing bond and use the resulting trajectory snapshots to compute time-averaged stresses at each atom. This would effectively sample multiple pathways for the dissociation process, leading to lower computed forces and stresses. Such a result would be more consistent with the lower forces observed in single-molecule AFM measurements for related host−guest systems.[45,46] It is also important to note that the bimolecular rupture (unbinding) force measured by AFM is not an intrinsic property of the molecular system, but instead depends upon the rate at which the two molecules are pulled apart, with higher rates leading to greater forces.[47,48,44] Thus, an even better, though more challenging, calculation would be to model the dynamic pulling process itself.

This paper also describes two methodological contributions to the application of stress calculations at the molecular level. One is the calculation of changes in stress in internal molecular coordinates, rather than in lab-frame coordinates, in order to enable meaningful comparisons of stresses among different conformational states and molecular orientations. The second is the method of displaying atomistic stress tensors. This is not trivial challenge, given that the stress on each atom is a symmetric 3×3 stress tensor with 6 distinct components.

The present study supports and advances the usefulness of computational stress analysis at the molecular level. It proves to be remarkably straightforward and informative to compute, compare, and visualize the stresses associated with various force components for host−guest systems. It will be interesting in the future to explore broader applications, especially to biomolecules, for which studies of static and dynamic stress can bear on atomistic mechanisms of conformational change that are of enormous scientific and practical importance.

Acknowledgment. I thank Dr. Yoshihisa Inoue for helpful discussions and Drs. Yoshihisa Inoue, Mikhail V. Rekharsky, and Cheng Yang for sharing the CB7-B5 affinity result. I also thank Dr. Hillary S. R. Gilson for helpful discussions and critical reading of the manuscript. This publication was made possible by grant no. GM61300 from the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

(1) Lutsko, J. *J. Appl. Phys.* **1988**, *64*, 1152–1154.

(2) Tsai, D. *J. Chem. Phys.* **1979**, *70*, 1375–1382.

(3) Cheung, K.; Yip, S. *J. Appl. Phys.* **1991**, *70*, 5688–5690.

(4) Zhou, M. *Proc. R. Soc. London A* , *459*, 2347–2392.

(5) Hardy, R. *J. Chem. Phys.* 1982.

(6) Zimmerman, J. A.; Webb, E. B., III; Hoyt, J. J.; Jones, R. E.; Klein, P. A.; Bammann, D. J. *Modelling Simul. Mater. Sci. Eng.* **2004**, *12*, S319–S332.

(7) Rafii-Tabar, H. *Phys. R* **2004**, *390*, 235–452.

(8) Koike, K.; Kawaguchi, K.; Yamato, T. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1400–1405.

(9) Yamato, T.; Higo, J.; Seno, Y.; Go, N. *Proteins* **1993**, *16*, 327–340.

(10) Buehler, M.; Ketena, S.; Ackbarow, T. *Prog. Mat. Sci.* **2008**, *53*, 1101–1241.

(11) Freeman, W.; Mock, W.; Shih, N. *J. Am. Chem. Soc.* **1981**, *103*, 7367–7368.

(12) Kim, J.; Jung, I.; Kim, S.; Lee, E.; Kang, J.; Sakamoto, S.; Yamaguchi, K.; Kim, K. *J. Am. Chem. Soc.* **2000**, *122*, 540–541.

(13) Hoffman, R.; Knoche, W.; Fenn, C.; Buschmann, H. *J. Chem. Soc. Farad. Trans.* **1994**, *90*, 1507–1511.

(14) Liu, S.; Rupic, C.; Mukhpadhayay, P.; Chakrabarti, S.; Zavalij, P. Y.; Isaacs, L. *J. Am. Chem. Soc.* **2005**, *127*, 15959–15967.

(15) Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L.; Chen, W.; Gilson, M. K.; Kim, K.; Inoue, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 20737–20742.

(16) Hwang, I.; Baek, K.; Jung, M.; Kim, Y.; Park, K.; Lee, D.; Selvapalam, N.; Kim, K. *J. Am. Chem. Soc.* **2007**, *129*, 4170–4171.

(17) Lagona, J.; Mukhopadhyay, P.; Chakrabarti, S.; Isaacs, L. *Angew. Chem., Int. Ed.* **2005**, *44*, 4844–4870.

(18) Kim, K.; Selvapalam, N.; Ko, Y. H.; Park, K. M.; Kim, D.; Kim, J. *Chem. Soc. Rev.* **2007**, *36*, 267–279.

(19) Isaacs, L. *Chem. Commun.* **2009**, *6*, 619–629.

(20) Tuncel, D.; Steinke, J. *Chem. Commun.* **1999**, *16*, 1509–1510.

(21) Park, K.; Heo, J.; Roh, S.; Jeon, Y.; Whang, D.; Kim, K. *Mol. Cryst. Liq. Cryst. Sci. Sect. A* **1999**, *327*, 65–70.

(22) Montes-Navajas, P.; Corma, A.; Garcia, H. *ChemPhysChem* **2008**, *9*, 713–720.

(23) Mock, W.; Irra, T.; Wepsiec, J.; Adhya, M. *J. Org. Chem.* **1989**, *54*, 5302–5308.

(24) Huang, W.; Liu, S.; Zavalij, P.; Isaacs, L. *J. Am. Chem. Soc.* **2006**, *128*, 14744–14745.

(25) Mock, W. L.; Shih, N. *J. Am. Chem. Soc.* **1989**, *111*, 2697–2699.

(26) Marquez, C.; Hudgins, R. R.; Nau, W. M. *J. Am. Chem. Soc.* **2004**, *126*, 5806–5816.

(27) Cram, D. J.; Tanner, M. E.; Knobler, C. B. *J. Am. Chem. Soc.* **1991**, *113*, 7717–7727.

(28) Pluth, M. D.; Raymond, K. N. *Chem. Soc. Rev.* **2007**, *36*, 161–171.

(29) Marquez, C.; Nau, W. M. *Angew. Chem., Int. Ed.* **2001**, *40*, 3155–3160.

(30) Moghaddam, S.; Inoue, Y.; Gilson, M. *J. Am. Chem. Soc.* **2009**, *131*, 4012–4021.

(31) Chang, C.; Gilson, M. K. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.

(32) Chen, W.; Chang, C.; Gilson, M. K. *Biophys. J.* **2004**, *87*, 3035–3049.

(33) Gilson, M. K.; Honig, B. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 5–20.

(34) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(35) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem.* **1997**, *101*, 3005–3014.

(36) Cohen, E.; Cvitas, T.; Frey, J.; Holmstrom, B.; Kuchitsu, K.; Marquardt, R.; Mills, I.; Pavese, F.; Quack, M.; Stohner, J.; Strauss, H.; Takami, M.; Thor, A. *Quantities, Units and Symbols in Physical Chemistry (the IUPAC Green Book)*, 3rd ed.; RSC Publishing: Cambridge, U. K., 2007.

(37) McNaught, A.; Wilkinson, A. *Compendium of Chemical Terminology (the "Gold Book")*, 2nd ed.; Blackwell Scientific Publications: Oxford, U. K., 1997.

(38) Hardy, R. *J. Chem. Phys.* **1982**, *76*, 622–628.

(39) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.

(40) Moore, J.; Schorn, S.; Moore, J. *J. Turbomachinery* **1996**, *118*, 622–629.

(41) Hashash, Y. M. A.; Yao, J. I.; Wotring, D. C. *Int. J. Numer. Analyt. Meth. Geomech.* **2003**, *27*, 603–636.

(42) Kindlmann, G.; Westin, C. *IEEE Trans. Visualiz. Comput. Graph.* **2006**, *12*, 1329–1335.

(43) Jankun-Kelly, T. J.; Mehta, K. *IEEE Trans. Visualiz. Comput. Graph.* **2006**, *12*, 1197–1204.

(44) Dudko, O. K.; Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 15755–15760.

(45) Kim, J.; Kim, Y.; Baek, K.; Ko, Y. H.; Kim, D.; Kim, K. *Tetrahedron* **2008**, *64*, 8389–8393.

(46) Yasuda, S.; Okutsu, Y.; Suzuki, I.; Shinohara, K.; Komiyama, M.; Takeuchi, O.; Shigekawa, H. *Jpn. J. Appl. Phys.* **2007**, *46*, 5614–5616.

(47) Lee, N.; Thirumalai, D. *Biophys. J.* **2004**, *86*, 2641–2649.

(48) West, D. K.; Olmsted, P. D.; Paci, E. *J. Chem. Phys.* **2006**, *124*, 154909.

CT900668K

JCTC Journal of Chemical Theory and Computation

The Exchange-Energy Density Functional Based on the Modified Becke-Roussel Model

Hideaki Takahashi,* Ryohei Kishi, and Masayoshi Nakano

Division of Chemical Engineering, Department of Materials Engineering Science, Graduate School of Engineering Science, Osaka University, Toyonaka Osaka 560-8531, Japan

Received August 10, 2009

Abstract: In this paper, we present a simple numerical approach to implement the modified Becke−Roussel (mBR) model for the purpose of developing an exchange density functional suitable for applications to atoms or molecules. Three steps constitute our approach. The first step is to model the exchange hole with the mBR distribution with the form of $\rho_{X\sigma}^{mBR} = (\alpha/\pi)^{3/2} \exp(-\alpha r^2)$ at each reference point, where α and r represent, respectively, the diffuseness and the distance of the model exchange hole from the reference point. We propose an iterative procedure to determine the values (α, r) during the Kohn−Sham DFT calculation. Second, we make a GGA correction to the functional obtained in the first step by adopting the conventional GGA formula to the gradients of the spin density as well as the mBR exchange hole (mBR-GGA). In the third step, mBR-GGA is combined with Dirac's exchange functional to restore the exchange energy at the homogeneous electron gas limit (mBR-hyb). We demonstrate that the exchange energy densities of the mBR-based methods obey the $-1/r$ asymptotic behaviors by virtue of the fact that the electron density in a hydrogenic atom is used as a prototypical exchange hole. Furthermore, we perform several test calculations for the properties of small molecules. For atomization energies for 35 molecules in the G2 set, the mean absolute deviation (MAD) with respect to the experiment is estimated to be 4.9 kcal/mol by the mBR-hyb functional, which is much smaller than the value of PBE functional (7.7 kcal/mol). The MAD for the enthalpies of formation of 68 molecules in the G3 set is evaluated as 9.4 kcal/mol by the present method, while that is given as 18.7 kcal/mol by the PBE functional. These results suggest the possibility of the present functional based on the mBR model for the applications to atoms or molecules.

1. Introduction

The density functional for the exact exchange energy of homogeneous electron gas was first formulated by Dirac in an attempt to establish a purely density-functional approximation for the electronic energy of a system.[1] The exchange functional was then introduced into the effective Hamiltonian in place of the exact exchange potential by Slater for the purpose of simplifying the Hartree−Fock method.[2] Later, this approach was reinterpreted within the rigorous framework of the Kohn−Sham density functional theory (KS-DFT)[3] and was validated as the local density approximation (LDA). The possibility of gradient corrections to the inhomogeneous electron density in a real system was first suggested by Hohenberg and Kohn[4] and afterward successfully taken into account by the method termed generalized gradient approximations (GGA).[5,6] The simplest form of the gradient correction is based on the lowest order gradient expansion,[7] which is, however, divergent in the limit of low electron density. In 1988, Becke proposed a nondivergent form of the functional satisfying correct asymptotic behavior of the exchange energy density (B88),[8] where one adjustable parameter was introduced in the correction term. Later in 1996, Perdew, Burke, and Ernzerhof developed an empirical-parameter-free exchange functional by imposing

* To whom correspondence should be addressed. E-mail: takahasi@cheng.es.osaka-u.ac.jp.

10.1021/ct900416x
Published on Web 02/01/2010

some physical constraints (PBE).[9] These exchange formulas in combination with correlation functionals achieved almost comparable accuracies to some molecular orbital theories considering the electron correlations. The success of the GGA correction encouraged subsequent improvements of the functional. In 1993, Becke proposed to mix the exact exchange potential to incorporate the kinetic correlation energy[10] in the functional through the adiabatic connection method. Perdew et al. made remarkable improvements in the computation of the molecular atomization energies by the method termed meta-GGA that includes the kinetic energy density and/or the Laplacian of the electron density in the functional.[11] A number of functionals are also being developed to compensate for other deficiencies inherent in the LDA-based KS-DFT such as the self-interaction error (SIE)[5,6,12,13] or the lack of long-range behavior of the exchange potential.[6,14,15] As shown in the Figure 1 in ref 16, each step in the development is often compared to various rungs of "Jacob's ladder" that may lead to the heaven of chemical accuracy.

Here, we pose a question whether the ladder in which LDA is regarded as the first rung is the only way to improve the exchange-correlation functionals. The exchange hole distribution in the homogeneous electron gas is spherically symmetric around the reference electron as a matter of course. In a bulk system, such hole behavior simulated by LDA gives a reasonable description for the real exchange hole. The bulk system has no boundary of the electron cloud, and hence, the exchange hole is always localized to some extent in the vicinity of the reference electron. On the contrary, in a finite molecular system, the exchange hole resides at the molecule even when the reference electron is placed far apart from the molecule. This is the origin of the fact that the exact exchange energy density as well as the exchange potential obeys $-1/r$ asymptotic behavior given that r is the distance between the reference electron and the molecule. Nevertheless, the hole depth within the LDA description decreases exponentially as r increases because of the exponential decay of the electron density. This results in a well-known unsatisfactory short-range behavior of the exchange energy density.[8] Thus, there is still room to consider another candidate for the exchange functional for the applications to atoms and molecules instead of the conventional route that begins with the homogeneous electron gas.

In this respect, Becke and Roussel (BR) proposed a unique exchange hole model for *inhomogeneous* systems.[17,18] They introduced the density of a 1s electronic wave function in a hydrogenic atom as a model of the exchange hole in molecular systems. The nuclear charge of the atom (which determines the diffuseness of the hole) and the distance of the hole from the reference electron are determined by imposing a condition that the model hole realizes the behavior of the spherically averaged exchange hole of the real system near the reference point. The exchange energy density is, then, computed as the Coulomb interaction between the reference electron and the exchange hole. The BR approach seems to be a natural choice for the exchange-hole model for the systems suffering serious inhomogeneity, such as atoms or molecules. It is worth noting that the long-range nature of the exchange energy density is fully restored in the BR model since it takes the hydrogenic atom as a prototypical system. In the subsequent developments by Becke et al., the BR model was utilized in the functional as a device to simulate the static correlations[19] or to generate the dispersion interactions.[20] The drawback of the BR approach is that the functional derivative with respect to the electron density cannot be evaluated explicitly, and hence, the variational potential appearing in the KS equation cannot be obtained as suggested in ref 18. In the writing of this paper, we noticed that Neumann et al. provided an efficient numerical technique to implement the BR functional into the self-consistent calculations in the KS-DFT with LCAO (linear combination of atomic orbitals) basis sets.[21,22] However, its realization in the computation with the real-space grids method[23] as well as the plane wave basis would still remain to be difficult. Although the BR model has many encouraging properties, so far there have been only a small number of numerical examples to the best of our knowledge.[18,22,24] In addition, in most cases in these studies, the one-electron wave functions were given at the outset, and the BR model was only used to evaluate the exchange energies for the given electron densities. In the present work, to quest for a new route to the exchange functional for practical applications, we propose a simple numerical approach efficient for any choice of the basis set within the framework of the BR model. Our strategy to implement the BR approach is to express the nuclear charge of the hydrogenic atom (or equivalently the exponent of the hole) as a functional of the electron density at the reference point. Then, the distance between the reference point and the hole is determined by ensuring that the hole density coincides with the electron density at the reference point. Second, the gradient correction to this scheme is also taken into consideration by utilizing the conventional GGA formula.

The organization of this paper is as follows. Section 2 is devoted to describing the details of the methodology. We first present a concise review for the original BR approach. Then, our approach is presented for the numerical implementation of the modified BR model in KS-DFT. The computational details for the test calculations utilizing the real-space grids approach are presented in section 3. We examine in section 4 the behaviors of the exchange energy density as well as the exchange potential with respect to the variation of the distance between the reference point and a molecule of interest. We also study the efficiency of the method by computing several properties of the small molecules such as atomization energy, ionization potentials, etc. In section 5, we provide a summary and conclusions to discuss the possibilities of the BR-model-based exchange functional.

2. Methodology

2.1. Exchange Hole Based on the Becke−Roussel Model. In the Becke−Roussel (BR) approach,[18] the exchange hole distribution $\rho_{X\sigma}(R)$ for spin σ in a molecule is

represented by a Slater function, which is a normalized 1s orbital density of a hydrogenic atom:

$$\rho_{X\sigma}^{BR}(R) = \frac{\alpha^3}{8\pi}\exp(-\alpha R) \quad (1)$$

In eq 1, R is the distance of the center of the hole from the reference electron. α specifies the diffuseness of the hole and is related to the nuclear charge of the hydrogenic atom. Provided that α and R are known for any given reference point $\mathbf{r}$, the exchange energy density $U_{X\sigma}^{BR}(\mathbf{r})$ based on the BR approach is expressed by the Coulomb interaction between the reference electron and the exchange hole distribution; thus,

$$U_{X\sigma}^{BR}(\mathbf{r}) = -\int_0^\infty ds \frac{1}{s}\int_{\Omega_s} \rho_{X\sigma}^{BR}(\mathbf{r}+\mathbf{s})\, d\mathbf{s} \quad (2)$$

In eq 2, Ω_s denotes the integration over a sphere of radius s centered at the reference point $\mathbf{r}$. The energy of eq 2 represents the exchange energy per electron in the BR model at the point $\mathbf{r}$ and is often referred to as exchange energy density. Note that $U_{X\sigma}^{BR}(\mathbf{r})$ is only dependent on the values (α, R). The exchange energy for spin σ is, then, given by using $U_{X\sigma}^{BR}(\mathbf{r})$; thus,

$$E_{X\sigma}^{BR} = \frac{1}{2}\int d\mathbf{r}\, \rho_{X\sigma}(\mathbf{r})\, U_{X\sigma}^{BR}(\mathbf{r}) \quad (3)$$

For the evaluation of the exchange energy, it is required to obtain the values (α, R) in eq 1 at every reference point. Becke and Roussel proposed to use these values as parameters to mimic the realistic exchange hole by a proper fitting at each reference point. Explicitly, the values (α, R) are determined by imposing the condition that the spherical average of the Taylor expansion of the model exchange hole around the reference point reproduces that of the real system up to the second order term of the expansion. In the following, we briefly review the procedure to determine (α, R). The exact spherically averaged exchange hole near the reference point $\mathbf{r}$ is expressed by

$$\rho_{SA-X\sigma}(\mathbf{r}, s) = \rho_\sigma(\mathbf{r}) + \frac{1}{6}(\nabla^2\rho_\sigma - 2\gamma D_\sigma)s^2 + \cdots \quad (4)$$

where s is the radius of the sphere centered at $\mathbf{r}$. $\rho_\sigma(\mathbf{r})$ in eq 4 is the electron density with σ spin, and D_σ is given by

$$D_\sigma = \tau_\sigma - \frac{1}{4}\frac{(\nabla\rho_\sigma)^2}{\rho_\sigma} \quad (5)$$

where τ_σ is the kinetic energy density and is defined as

$$\tau_\sigma = \sum_i |\nabla\phi_{\sigma i}|^2 \quad (6)$$

γ in eq 4 is the parameter to be used in later reference, and the equality $\gamma = 1$ holds in the exact expression.

For the exchange hole model defined by eq 1, the spherically averaged exchange hole is analytically expressed by a function of the variables (α, R) as

$$\rho_{SA-X\sigma}^{BR}(\alpha, R; s) = \frac{\alpha}{16\pi Rs}[(\alpha|R-s|+1)\exp(-\alpha|R-s|) - (\alpha|R+s|+1)\exp(-\alpha|R+s|)] \quad (7)$$

By equating the first two coefficients of the Taylor expansion of eq 7 to those of the real system given by eq 4, we obtain two equations:

$$\alpha^3\exp(-\alpha R) = 8\pi\rho_\sigma \quad (8)$$

and

$$\alpha^2 R - 2\alpha = 6RQ_\sigma/\rho_\sigma, Q_\sigma = \frac{1}{6}(\nabla^2\rho_\sigma - 2\gamma D_\sigma) \quad (9)$$

Then, they lead to the following equation with the definition of $x = \alpha R$,

$$\frac{x\exp(-2x/3)}{x-2} = \frac{2}{3}\pi^{2/3}\frac{\rho_\sigma^{5/3}}{Q_\sigma} \quad (10)$$

The variable x in eq 10 is obtained by using the Newton−Raphson scheme, from which the parameters (α, R) are determined. Importantly, eq 10 assures the existence of a unique and positive root x for all conditions. The exchange energy density at reference point $\mathbf{r}$ is, then, given by

$$\begin{aligned} U_{X\sigma}^{BR}(\mathbf{r}) &= -4\pi\int_0^\infty ds\, \rho_{SA-X\sigma}^{BR}(\alpha, R; s)s \\ &= -\left(1 - \exp(-R) - \frac{1}{2}R\exp(-R)\right)/R \end{aligned} \quad (11)$$

It is readily verified that $U_{X\sigma}^{BR}(\mathbf{r})$ behaves as $-1/r$ when $\mathbf{r}$ is placed far apart from the molecular system. It is worth noting that one of the important features to be fulfilled by the approximate exchange functional is, thus, naturally incorporated in the BR functional from the outset by virtue of the fact that the atomic electron density is chosen as a prototypical system.

2.2. Modified Becke-Roussel Model and Its Implementation. Here, we introduce a modified Becke−Roussel (mBR) model where the exchange hole is modeled by a Gaussian function instead of the Slater type one as given by eq 1. Recently, such a modification was also suggested by Bahmann and Ernzerhof,[25] who mixed the mBR model with LDA by a switching factor, where the missing cusp of the hole distribution is traded for the simplified integrals. In the present work, we introduce the mBR model for the purpose of suiting the BR model to the Kohn−Sham equation that utilizes pseudopotentials. The exact exchange hole at a given reference point is described in terms of the one-electron wave functions, and hence, the exchange-hole distribution reflects the behavior of the one-electron wave functions to some extent. Since the point of the BR approach is to express the spherical average of the hole distribution by that of the atomic orbital density, a Gaussian function will be well suited to mimic the behavior of the pseudo-wave functions of which variation near the nuclei are forced to be sufficiently smooth. On the contrary, the Slater function of eq 1 is appropriate for modeling the exchange hole in all-electron calculations where the wave functions have cusps

near the nuclei. In the present work, we propose a numerical approach for implementing both the mBR and the original BR methods. Within the mBR model, the exchange hole is represented by a normalized Gaussian function:

$$\rho_{X\sigma}^{mBR}(R) = \left(\frac{\alpha}{\pi}\right)^{3/2}\exp(-\alpha R^2) \tag{12}$$

The notations in eq 12 are synonymous with those in eq 1. As discussed for the original BR model in section 2, the determination of the parameters (α, R) is also crucial in the mBR approach. We propose here a simpler numerical approach to obtain these values than those proposed in refs 18 and 21. At first, we consider two limiting situations for a reference electron. In case 1, the reference electron is assumed to be placed at the position where the electron density of the system has its maximum value ρ_σ^{max} ($\rho(\mathbf{r}) = \rho_\sigma^{max}$). And in case 2, it is assumed that the reference point is placed far apart from the molecular system ($\rho(\mathbf{r}) \cong 0$). The concept of the Fermi orbital suggests that the behavior of the exchange hole is dominated by the orbital that gives a major contribution to the total density at the reference point. Hence, it is reasonable to consider that the exchange-hole distributions for the reference electrons in cases 1 and 2 are characterized by the core orbital and the HOMO of the system, respectively. Then, for these two limiting situations, it is possible to estimate approximately the parameters (α, R) in eq 12. For case 1, the center of the exchange hole may coincide well with the position of the reference electron, which implies that the variable R in eq 12 can be taken as zero. Further, we impose a physical constraint that the exchange-hole density at the reference point is exactly the same with the electron density

$$\rho_{X\sigma}^{mBR}(R)|_{R=0} = \rho_\sigma^{max} \tag{13}$$

This condition can be easily verified by taking $s = 0$ in the expansion of eq 4. Then, we have

$$\alpha_1 = \pi\rho_\sigma^{max\ 2/3} \tag{14}$$

where α_1 denotes specifically the value of α for case 1. As for the reference electron in the opposite situation (case 2), the exchange hole obeys the asymptotic form as derived in refs 26 and 27; thus,

$$\lim_{R\to\infty} \rho_{X\sigma} = \exp(-\alpha'_0 R) \tag{15}$$

Equation 15 is consistent with the fact that the wave function decays exponentially in the asymptotic region. More importantly, the exponent α_0' in eq 15 can be approximately related to the ionization potential I as follows:

$$\frac{1}{2}\alpha'^2_0 = I \tag{16}$$

According to the proof given in ref 28, the eigenvalue ϵ_{HOMO} of the exact Kohn−Sham equation is identical to $-I$. Hence, α_0' in eq 16 can be approximated as

$$\alpha'_0 = (-2\varepsilon_{HOMO})^{1/2} \tag{17}$$

To employ the mBR model for describing the asymptotic hole distribution, the Slater function with exponent α_0' given by eq 17 has to be fitted by a Gaussian function. As shown in ref 29, the value α_0, defined as the exponent for case 2, can be derived from α_0' by least-squares fitting and the scaling relation; thus,

$$\alpha_0 = \alpha_0^{1.0} \times \alpha'^2_0 \tag{18}$$

where $\alpha_0^{1.0}$ is the exponent of the Gaussian function fitted to the Slater one with the exponent $\alpha_0' = 1.0$. Thus, we can deduce approximately the exponents α_1 and α_0 with $R = 0$ and $R \to \infty$ corresponding to the opposite situations $\rho(\mathbf{r}) = \rho_\sigma^{max}$ and $\rho(\mathbf{r}) = 0$, respectively. Here, we should note that the use of ϵ_{HOMO} leads to an undesirable consequence that the size consistency cannot be fulfilled. To show this, we consider a complex of two monomers with different HOMO energies at a large separation. Then, the HOMO energy of the complex will be the same as the constituent molecule with the larger eigenvalue. Hence, the complex does not have the same energy as the separated fragments. Thus, our present approach violates the size consistency. This limitation should be kept in mind in simulating the dissociation processes.

For the determination of the exponent α in eq 12 for the intermediate reference point $\mathbf{r}$, which satisfies $0 < \rho_\sigma(\mathbf{r}) < \rho_\sigma^{max}$, we introduce the interpolation

$$\frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0} = \left(\frac{\rho_\sigma(\mathbf{r})}{\rho_\sigma^{max}}\right)^p \tag{19}$$

where p is the scaling parameter. Once the exponent α is, thus, obtained, the distance R between the reference point and the exchange hole can be simply derived from the relation $\rho_\sigma(\mathbf{r}) = \rho_{X\sigma}^{mBR}(R)$ as

$$R = \left(-\frac{1}{\alpha}\log\left(\left(\frac{\pi}{\alpha}\right)^{3/2}\rho_\sigma(\mathbf{r})\right)\right)^{1/2} \tag{20}$$

Since the exchange energy density $U_{X\sigma}^{mBR}$ is the Coulomb interaction between the reference electron and the exchange hole, $U_{X\sigma}^{mBR}$ is simply given by

$$U_{X\sigma}^{mBR}(\mathbf{r}) = -\frac{1}{R}\mathrm{Erf}(-\alpha^{1/2}R) \tag{21}$$

with the definitions of eqs 19 and 20. Then, the exchange energy $E_{X\sigma}^{mBR}$ of the system based on the mBR model is given by

$$E_{X\sigma}^{mBR} = \frac{1}{2}\int d\mathbf{r}\ \rho_\sigma(\mathbf{r})\ U_{X\sigma}^{mBR}(\mathbf{r}) \tag{22}$$

It should be noted that the functional of eq 22 fulfills the important property that the exchange hole contains just one electron. Furthermore, it is ensured by the construction that the exchange-hole density at the reference point is exactly equal to the electron density (i.e., the first term of the expansion of eq 4 is ensured). Here, we also note that it can be readily proved there exists no real value of R for $p > 2/3$ in eq 19. In other words, we can choose p as a scaling parameter to control the diffuseness of the exchange hole with respect to the magnitude of the spin density. At first,

we check the performance of the functional with the critical value of $p = 2/3$ to assess the efficiency of the present approach based on the BR model and then optimize the parameter to refine the functional (see details in section 2.4).

For the implementation of the mBR approach in the Kohn−Sham equation, it is necessary that the exchange potential $v_{X\sigma}^{mBR}(\mathbf{r})$ is obtained by the derivative of $E_{X\sigma}^{mBR}$ with respect to the density $\rho_\sigma(\mathbf{r})$,

$$v_{X\sigma}^{mBR}(\mathbf{r}) \equiv \frac{\delta E_{X\sigma}^{mBR}(\mathbf{r})}{\delta \rho_\sigma} = \frac{1}{2}\left(U_{X\sigma}^{mBR}(\mathbf{r}) + \rho_\sigma(\mathbf{r}) \frac{\partial U_{X\sigma}^{mBR}(\mathbf{r})}{\partial \rho_\sigma} \right) \tag{23}$$

As described in the previous paragraph, the exchange functional $E_{X\sigma}^{mBR}$ defined by eqs 19−22 contains the parameters α_1 and α_0. We note that these parameters are also dependent on the electron density; however, the derivatives of the exponents α_1 and α_0 with respect to $\rho_\sigma(\mathbf{r})$ are unavailable. Here, we adopt the following numerical approach to determine these parameters iteratively. Analytical expression $\tilde{v}_{X\sigma}^{mBR}(\mathbf{r})$ for the derivative $\delta U_{X\sigma}^{mBR}(\mathbf{r})/\delta\rho_\sigma$ can be readily obtained by supposing that the values α_1 and α_0 are fixed. Then, we consider solving the following Kohn−Sham equation with the exchange potential $v_{X\sigma}^{mBR\prime}$,

$$\left[-\frac{1}{2}\nabla^2 + V_{H-ps}(\mathbf{r}) + \tilde{v}_{X\sigma}^{mBR}(\mathbf{r}) + v_{c\sigma}(\mathbf{r}) \right] \varphi_{i\sigma}(\mathbf{r}) = \varepsilon_i \varphi_{i\sigma}(\mathbf{r}) \tag{24}$$

where V_{H-ps} is the sum of the Hartree and pseudopotentials, and $v_{c\sigma}$ denotes the correlation potential. The solution of eq 24 leads to the electron density $\rho_\sigma(\mathbf{r})$ and ϵ_{HOMO}, from which we derive new values of α_1 and α_0 by eqs 14, 17, and 18. Using the renewed α_1 and α_0, we reconstruct the potential $\tilde{v}_{X\sigma}^{mBR}(\mathbf{r})$ and solve eq 24. This procedure is continued until the parameters α_1 and α_0 converge. We confirm in practice that the renewal of these values at every SCF step is sufficient for the convergence. This iterative approach will give rise to an instability in the SCF procedure to some extent; however, our test calculation shows a satisfying convergence rate as demonstrated in section 4.1. Thus, we obtain a simplified procedure based on the BR model for describing the exchange energy. In closing this paragraph, it should be noted that we are not guaranteed to have the proper exchange potential even at the convergence since it is not possible to obtain the correct Kohn−Sham eigenvalues.

So far, we have discussed the practical implementation of the mBR approach that is efficient for the calculations with the plane-wave basis and real-space grid methods utilizing pseudopotentials. Here, we also illustrate a method along this line for the original BR model with the form of eq 1. As for the exponent α_1' of the Slater-type exchange hole for the situation of case 1, we have

$$\alpha'_1 = 2(\pi\rho_\sigma^{max})^{1/3} \tag{14'}$$

as the counterpart of eq 14 for the Gaussian function. The exponent α_0' for case 2 is directly determined by eq 17 when we employ the Slater function. Then, the exponent α' for the intermediate reference point can be determined as

$$\frac{\alpha' - \alpha'_0}{\alpha'_1 - \alpha'_0} = \left(\frac{\rho(\mathbf{r})}{\rho_{max}} \right)^{p'} \tag{19'}$$

We note that the scaling parameter p' in eq 19′ can also be considered as an adjustable parameter which should satisfy $0 \le p' \le 1/3$. In accord with the choice of $p = 2/3$ in eq 19 for the case of the Gaussian-type exchange hole, p' is taken as 1/3 for eq 19′. Then, the distance r' between the reference point and the center of the Slater-type hole is expressed as

$$R' = -\frac{1}{\alpha'} \log\left(\frac{8\pi}{\alpha'^3} \rho_\sigma(\mathbf{r}) \right) \tag{20'}$$

The exchange energy density $U_{X\sigma}^{BR}$ is exactly expressed in the form given by eq 11. The subsequent procedure for the SCF calculation for the solution of the Kohn−Sham equation is essentially parallel to that for the mBR model.

2.3. Gradient Correction to Modified Becke−Roussel Model. The exchange functional introduced in the previous section is originated from the BR model and has an important property that the exchange energy density obeys $-1/r$ asymptotic behavior for the reference electron placed far apart from the molecule. It is obvious that this advantage is attributed to the nature of the atomic electron density chosen as a model exchange hole. However, our simplified approach does not involve the information of the gradient of the electron density in its functional. And, hence, it gives exactly the same value of the exchange energy density at two different points as far as they have the same electron density. Here, we propose a method to incorporate the gradient correction into the mBR approach (mBR-GGA) by utilizing the conventional GGA formalism. In the following, we present the formulation for the mBR-GGA approach. At first, we consider the expansion of the spherically averaged exchange hole for the mBR model that is completely parallel to eq 4; thus,

$$\rho_{SA-X\sigma}^{mBR}(\mathbf{r}, s) = \rho_{X\sigma}^{mBR}(\mathbf{r}) + \frac{1}{6}(\nabla^2 \rho_{X\sigma}^{mBR} - 2\gamma D_\sigma^{mBR}) s^2 + \cdots \tag{25}$$

The notational conventions in eq 25 are, of course, common with those in eq 4, except that the superscripts mBR are attached to each function. The exact spherically averaged exchange hole $\rho_{SA-X\sigma}(\mathbf{r},s)$ can be formally written as

$$\rho_{SA-X\sigma}(\mathbf{r}, s) = \rho_{SA-X\sigma}^{mBR}(\mathbf{r}, s) + (\rho_{SA-X\sigma}(\mathbf{r}, s) - \rho_{SA-X\sigma}^{mBR}(\mathbf{r}, s)) \tag{26}$$

By substituting the expansions given by eqs 4 and 25 into eq 26, we obtain an exact expression for $\rho_{SA-X\sigma}(\mathbf{r},s)$ as

$$\rho_{SA-X\sigma}(\mathbf{r}, s) = \rho_{SA-X\sigma}^{mBR}(\mathbf{r}, s) + \left(\left(\frac{1}{6}(\nabla^2 \rho_\sigma - 2\gamma D_\sigma) s^2 + \ldots \right) - \left(\frac{1}{6}(\nabla^2 \rho_{X\sigma}^{mBR} - 2\gamma D_\sigma^{mBR}) s^2 + \ldots \right) \right) \tag{27}$$

The relation of $\rho_{X\sigma}^{mBR}(\mathbf{r}) = \rho_\sigma(\mathbf{r})$ is used in taking the subtraction in the parentheses of eq 26. The exchange energy contribution due to the term $\rho_{SA-X\sigma}^{mBR}(\mathbf{r},s)$ in the right-hand side of eq 27 is what we discussed in section 2.2 and explicitly given by eq 22. The s^2 terms in the parentheses of eq 27 are the leading terms of the gradient corrections for

the electron density of the real system and the exchange hole distribution of the mBR model. Here, it should be reminded that the GGA approach is constructed to evaluate approximately the contribution due to the s^2 term in the gradient expansion. Our method to evaluate eq 27 is to compute gradient corrections in the parentheses by employing the available GGA exchange functionals $E_{X\sigma}^{GGA}$ such as B88 and PBE; thus,

$$E_{X\sigma}^{mBR\text{-}GGA} = E_{X\sigma}^{mBR} + (E_{X\sigma}^{GGA}[\rho_\sigma, \nabla\rho_\sigma] - E_{X\sigma}^{GGA}[\rho_\sigma^{mBR}, \nabla\rho_\sigma^{mBR}]) \quad (28)$$

In eq 28, the exchange energy contribution $E_{X\sigma}^{LDA}$ due to the homogeneous electron gas, which is included in the GGA functional, completely vanishes by the subtraction in the parentheses. Furthermore, when $|\nabla\rho_\sigma|$ is equal to $|\nabla\rho_{X\sigma}^{mBR}|$ at a reference point, the exchange energy density $U_{X\sigma}^{mBR\text{-}GGA}$ is the same with $U_{X\sigma}^{mBR}$ at that point due to the cancellation of the gradient terms. It is worth noting that no empirical parameter is newly introduced in the construction of eq 28. For the SCF procedure to solve the Kohn−Sham equation, the functional derivative of eq 28 with respect to density ($\nu_{X\sigma}^{mBR\text{-}GGA}(\mathbf{r}) \equiv \delta E_{X\sigma}^{mBR\text{-}GGA}/\delta\rho_\sigma$) is necessary. We take essentially the same procedure as that proposed in the previous section, that is, the exchange potential $\tilde{v}_{X\sigma}^{mBR\text{-}GGA}(\mathbf{r})$ in Kohn−Sham equation is computed with the exponents α_0 and α_1 frozen. Then, these values are to be renewed after every SCF cycle to construct a new exchange potential. This process is iterated until the exponents as well as the electron density are converged.

2.4. Combination of Modified Becke−Roussel Model with LDA. The exchange functionals $E_{X\sigma}^{mBR}$ and $E_{X\sigma}^{mBR\text{-}GGA}$ discussed so far are along a unique line that starts from an inhomogeneous electron density as a prototypical hole distribution. However, these functionals have a crucial deficiency that they do not yield the exact LDA exchange energy at the homogeneous electron gas limit. This situation is also true for the original BR model described in section 2.1. Becke and Roussel proposed a method to recover the LDA exchange energy at the homogeneous limit by substituting $\gamma = 0.8$ in eq 9 instead of the true value of 1.0. Here, we take a mixing scheme (mBR-hyb) which hybridizes the $E_{X\sigma}^{mBR\text{-}GGA}$ with an LDA based exchange functional in a similar way to that proposed by Bahmann and Ernzerhof.[25] That is, we introduce the hybrid exchange functional $E_{X\sigma}^{mBR\text{-}hyb}$ as

$$E_{X\sigma}^{mBR\text{-}hyb} = \frac{1}{2}\left(\int d\mathbf{r}\, \exp(-cR^2)\rho_\sigma U_{X\sigma}^{LDA\text{-}GGA} + \int d\mathbf{r}\, (1 - \exp(-cR^2))\rho_\sigma U_{X\sigma}^{mBR\text{-}GGA}\right) \quad (29)$$

In eq 29, R is the distance between the reference point $\mathbf{r}$ and the center of the exchange hole described by the mBR model, and it is explicitly given by eq 20. $U_{X\sigma}^{LDA\text{-}GGA}$ denotes the standard LDA-based GGA exchange functional, and c expresses the mixing parameter. We note that, for the system with uniform electron density, $\rho_\sigma(\mathbf{r}) = \rho_\sigma^{max}$ holds wherever the reference point is chosen, and hence, R is equal to zero and the second term of the right-hand side of eq 29 vanishes. Thus, it is readily recognized that the functional of eq 29 recovers the exact exchange energy at the uniform gas limit. The mixing parameter c in eq 29 is, of course, an unknown parameter and considered to be an adjustable parameter to tune the distance at which the mBR exchange hole is mixing in. As noted in section 2.2, we at first take the parameter p in eq 19 as 2/3, and correspondingly the mixing parameter is chosen as $c = 1.0$ to ensure almost the exact value of 0.5 au for the electronic energy of a hydrogen atom. Then, the two-dimensional optimization for the parameter set (p, c) is carried out to attain the best performance of the present functional. It should be stressed that p and c are adjustable but surely contain physical meanings. Lastly, it should be noted that the GGA correction to the original BR model with the form of eq 1 can also be formulated in parallel to that for the mBR model described above.

3. Computational Details

Here, we present the computational details for the test calculations performed for small molecules. We have implemented the series of the mBR-based exchange functionals described above in our original code, which utilizes the real-space grids (RSG) and the pseudopotentials.[30−32] The methodological details for our RSG approach were presented in refs 23, 33, and 34. The kinetic energy operator in the one-electron Hamiltonian has been represented by the fourth-order finite-difference method. The nonperiodic hartree potential has been constructed by the method proposed in ref 35. The pseudopotentials derived by the method of Kleinmann and Bylander[36] have been used to express the effective potentials for valence electrons. A molecule of interest has been placed in the center of a cubic cell of which the axis has been uniformly discretized by 64 grids along each direction. The time-saving double grid approach proposed by Ono and Hirose[37] has been utilized to realize the rapid behavior of the nonlocal pseudopotentials as well as the pseudo-wave functions near the atomic cores. We have set the width of the original coarse grid at $h = 0.1518$ Å and that of the double grid at $h/3$. The convergence of the SCF procedure has been judged by the root-mean-square δ for the deviation of the electron density. To be specific, $\delta < 10^{-5}$ is imposed on the SCF convergence, which typically ensures the convergence within $10^{-6} \sim 10^{-7}$ E_h in the total electronic energy. The use of the pseudopotentials leads to an error due to its approximate construction. We have carefully checked the effects by comparing the atomization energies obtained by our code with those given by all-electron calculations with the Gaussian 03[38] program package. For these calculations, the PBE exchange has been used in combination with the Lee−Yang−Parr (LYP)[39] correlation functional. In the calculations by Gaussian 03, sufficiently large LCAO basis sets have been employed, where the quadruply split valence orbitals are used and polarization as well as diffuse atomic orbitals are augmented (aug-cc-pVQZ). Furthermore, the "Grid = UltraFine" option has been invoked to ensure the accuracy in the numerical integration of the DFT calculation.

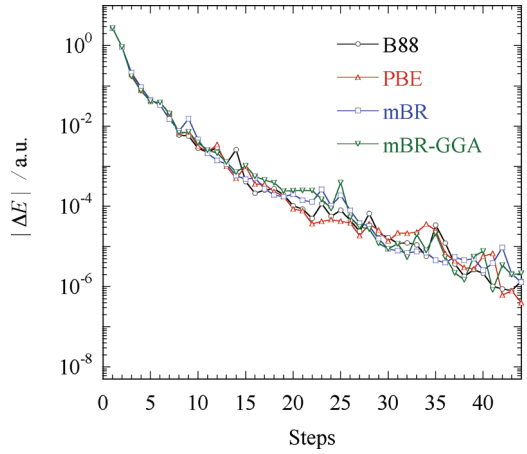


Figure 1. Convergence behaviors of the SCF procedures in KS-DFT that utilize the exchange functionals, B88, PBE, mBR, and mBR-GGA. Absolute values of the energy differences between neighboring steps are plotted against the SCF steps.

4. Applications and Tests

In this section, we present the results of the numerical applications of the mBR-based functionals and compare them with the conventional GGA functionals. In section 4.1, the convergence behaviors in the SCF procedures for these functionals are examined. In sections 4.2 and 4.3, the enhancement factors as well as the exchange energy densities and the exchange potentials have been plotted as functions of the position of the reference electron placed in a molecular system. In section 4.4, we present the results of the atomization energies, ionization potentials, proton affinities, and enthalpies of formation to discuss the accuracy and efficiency of the present approach. The exchange-hole functions given by this model for several reference points in a molecule are plotted in section 4.5 to make comparisons with LDA and the exact exchange holes. Hereafter, we refer to the exchange functionals defined by eqs 22, 28, and 29 by the shorthand notations mBR, mBR-GGA, and mBR-hyb, respectively.

4.1. Convergence Behavior. As described in section 2.2, the exchange energy functional $E_{X\sigma}^{mBR}$ of eq 22 includes density-dependent parameters (α, R) which are iteratively determined in the SCF procedure as well as the electron density. This possibly leads to a numerical instability in the SCF convergence to a certain extent. To see this in detail, we have examined the convergence rate in SCF for the mBR and mBR-GGA exchange functionals which are, respectively, given in the forms of eqs 22 and 28. The PBE exchange functional has been used for the GGA correction term $E_{X\sigma}^{GGA}$ in eq 28. We have also investigated the convergence behaviors of the SCF calculations with B88 and PBE exchange functionals to make comparisons. A water molecule has been used for these test calculations, where the wave functions have been updated by the scaled steepest descent (SD) algorithm starting from the same initial guess. The geometry of the water has been taken from ref 40.

In Figure 1, the log plots have been shown for the absolutes of the differences in the electronic energies between adjacent steps in the SCF procedure against the SCF step number. It is clearly shown that the convergence rates of the SCF calculations with mBR or mBR-GGA are comparable to those with the conventional GGA functionals such as B88 or PBE. Thus, it has been demonstrated that no serious numerical instability takes place in the SCF procedures for the practical applications of the series of the mBR exchange functional. We remind the reader that the pseudopotentials have been used through these test calculations, and they may possibly support the stability in the convergence. Unfortunately, we could not check the effect of the use of pseudopotentials since an LCAO-based program package equipped with our approach is not available.

4.2. Enhancement Factor. To investigate the property of an exchange functional mBR, we have evaluated the exchange enhancement factor $F_{X\sigma}^{mBR}$ defined by

$$E_{X\sigma}^{mBR} = \frac{1}{2}\int d\mathbf{r}\, \rho_\sigma \cdot U_{X\sigma}^{LDA} \cdot F_{X\sigma}^{mBR} \tag{30}$$

In eq 30, $U_{X\sigma}^{LDA}$ is the exchange energy density of the uniform electron gas derived by Dirac[1] and is given by

$$U_{X\sigma}^{LDA} = -3\left(\frac{3}{4\pi}\right)^{1/3}\rho_\sigma^{1/3} \tag{31}$$

Equations 22 and 30 suggest that the factor $F_{X\sigma}^{mBR}$ represents the ratio of the exchange energy density $U_{X\sigma}^{mBR}$ given in eq 21 with respect to $U_{X\sigma}^{LDA}$. We have performed the same evaluations for the functionals mBR-GGA and mBR-hyb. Usually, the enhancement factor for a GGA functional is plotted as a function of the dimensionless parameter $x_\sigma \equiv |\nabla\rho_\sigma|/\rho_\sigma^{4/3}$, which represents the inhomogeneity of the electron density since a usual GGA functional depends only on x_σ. However, our functional is also dependent on the maximum value of the electron density as well as the eigenvalue of the HOMO besides the gradient of the density. Therefore, we have computed $F_{x\sigma}$ for a real system with a closed shell electronic structure. Specifically, we have chosen a water molecule as a model system; then, the enhancement factors have been plotted by varying the position of the reference electron along the symmetry axis of the water molecule. For the calculations of the enhancement factors for each functional shown below, we have adopted the electron density obtained by using the B88 functional without correlation.

In Figure 2, we have plotted the enhancement factors for the functionals, mBR, mBR-GGA, and mBR-hyb. The PBE functional has been employed for the GGA correction in the mBR-GGA and mBR-hyb functionals. The factors for the functionals B88 and PBE have also been depicted in the figure to make comparisons. The inhomogeneity parameter x_σ has been superimposed in Figure 2. The horizontal axis of Figure 2 represents the coordinate of the reference electron placed on the symmetry axis of the water molecule for which an experimental geometry[40] has been used. The origin of the axis has been placed on the oxygen atom of the water. In addition, the center of molecular mass has been adjusted to the negative direction of the axis.

At first, we made comparisons between the behaviors of $F_{x\sigma}$ for the conventional GGA exchange functionals B88 and PBE. These two functionals start from the same basis;

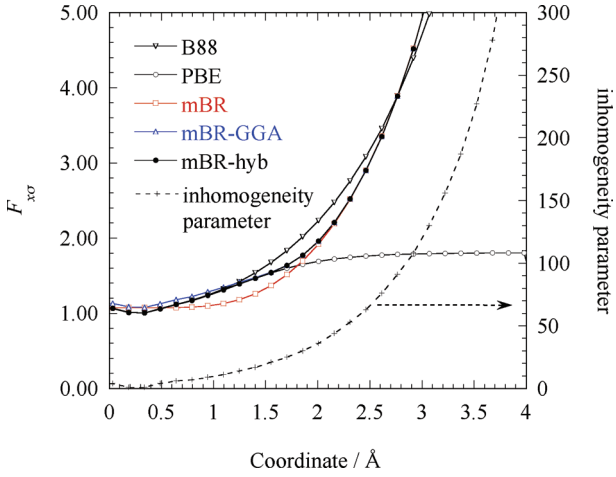


Figure 2. Plots for the enhancement factors given by B88, PBE, mBR, mBR-GGA, and mBR-hyb functionals. The horizontal axis represents the position of the reference electron placed on the symmetry axis of a water molecule. The oxygen atom is taken as the origin of the axis. The electron density obtained by the KS-DFT calculation using the B88 functional without correlation is employed to evaluate the enhancement factors with these functionals. The dimensionless parameter x_σ is also drawn in the figure.

nevertheless, $F_{x\sigma}$ for B88 shows a distinct deviation from that for PBE. We observe in Figure 2 the enhancement factor $F_{X\sigma}^{B88}$ for the B88 functional increases rapidly as the reference electron moves away from the molecule, and this behavior agrees with that of the inhomogeneity parameter x_σ. In contrast to B88, the enhancement factor $F_{X\sigma}^{PBE}$ for PBE increases modestly and converges to a certain value as r increases. Such a discrepancy in the region of large inhomogeneity can be attributed to the difference in the physical constraints that are imposed on these functionals. Both the B88 and the PBE exchange functionals originate from basically the same analytic function with a nondivergent property for the increase of x_σ.[9,41] On the basis of this function, the B88 functional was constructed so that it recovers the $-1/r$ asymptotic behavior of the exchange energy density,[8] while the PBE functional was subjected to the constraint of the local Lieb−Oxford bound.[42] Consequently, the enhancement factor for B88 increases rapidly in the asymptotic region, and that for PBE converges to 1.804 at the limit of the large inhomogeneity, as shown in Figure 2. Later, Zhang and Yang proposed a method termed revPBE by choosing the asymptotic value of $F_{X\sigma}^{PBE}$ as 2.245 instead of 1.804.[43] This modification can be validated by the fact that the adoption of the Lieb−Oxford bound for all positions of the reference electron is just a sufficient but not necessary condition to satisfy the integrated Lieb−Oxford bound. Anyway, we observe that the enhancement factors for B88 and PBE functionals coincide well in the region of small x_σ, and they diverge rapidly as the reference electron moves away from the molecule. From a numerical point of view, we note that such a divergence does not have serious effects on the energetics because the region of large inhomogeneity coincides with the small density tail, and therefore with exponentially small energy density owing to the $\rho^{1/3}$ term from the LDA part.

The enhancement factor $F_{X\sigma}^{mBR}$ for the functional mBR shows the correct asymptotic behavior similar to the B88 functional by virtue of the fact that the atomic electron density is adopted as a model exchange hole. Here, it should be emphasized that such an important property is built-in in the model and is naturally simulated without making a special device for it. In Figure 2, it can also be recognized that the factor $F_{X\sigma}^{mBR}$ is almost comparable to $F_{X\sigma}^{LDA}$ from the short to middle range of the reference position (note that $F_{X\sigma}^{LDA} = 1$ holds everywhere by the definition of eq 30). Thus, the deviation of the mBR functional from B88 or PBE has been found to be serious for the reference electron placed on the coordinate of ∼1.5 Å. The enhancement factor $F_{X\sigma}^{mBR\text{-}GGA}$ for eq 28 shows that this unpleasant situation can be substantially alleviated by the GGA correction adopted in the mBR functional.

As for the short-range behaviors of $F_{X\sigma}^{mBR}$ and $F_{X\sigma}^{mBR\text{-}GGA}$, it is shown in Figure 2 that they slightly overestimate those of the LDA-based exchange functionals. According to the prescription by Perdew et al.,[16] an exchange functional should reproduce Dirac's exact exchange at the uniform electron gas limit. The functional mBR-hyb given by eq 29 is designed so that it recovers the LDA exchange energy at the uniform gas limit. Consequently, the enhancement factor $F_{X\sigma}^{mBR\text{-}hyb}$ realizes the sound behavior in the region of the small inhomogeneity. Thus, the mBR-hyb functional has a desirable property in the enhancement factor from the short- to long-range region of the reference electron.

4.3. Exchange Energy Density and Exchange Potential. One of the important features of the mBR-model based approach lies in the possibility that it can recover the correct long-range behaviors of the exchange energy density $U_{x\sigma}$ and the exchange potential $v_{x\sigma}$. $U_{x\sigma}$ represents the exchange energy per electron felt at a given position and is explicitly defined by an equation parallel to eq 3. It can be readily verified that exact $U_{x\sigma}$ behaves as $-1/r$ in the asymptotic region of $r \rightarrow \infty$ (r is roughly the distance between the reference point and the molecule of interest). The exchange potential $v_{x\sigma}$ given by the functional derivative of exchange energy $E_{x\sigma}$ with respect to density (parallel to eq 23) is also known to decay asymptotically in the Coulombic manner as $-1/r$. In the following, we have investigated the behaviors of $U_{x\sigma}$ and $v_{x\sigma}$ obtained with the exchange functional mBR-hyb and compared it with conventional GGA functionals. For these calculations, we have employed the same system used in the calculation of the enhancement factors presented in section 4.2.

In Figure 3, the exchange energy density for the mBR-hyb functional has been plotted for the variation of the position of the reference point. Those for the B88 and the PBE functionals have also been drawn in the figure. The horizontal axis is the coordinate of the reference electron of which the origin has been taken in the same manner as in section 4.2. For the construction of the Coulomb potential $-1/r$ in the figure, the origin has been set at the nuclear-charge weighted center of mass. Explicitly, a coordinate of −0.118 Å is chosen as the origin of the Coulomb potential. We observe in Figure 3 that the exchange energy density for B88 functional

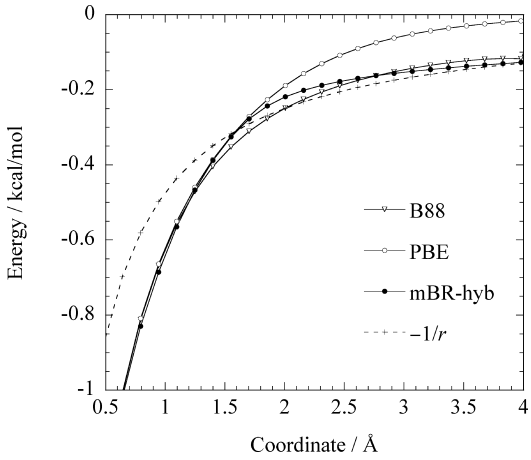


Figure 3. Plots for the exchange energy densities given by B88, PBE, and mBR-hyb functionals. The definition of the horizontal axis is the same as in Figure 2. The electron density used in the construction of Figure 2 is employed for these calculations.

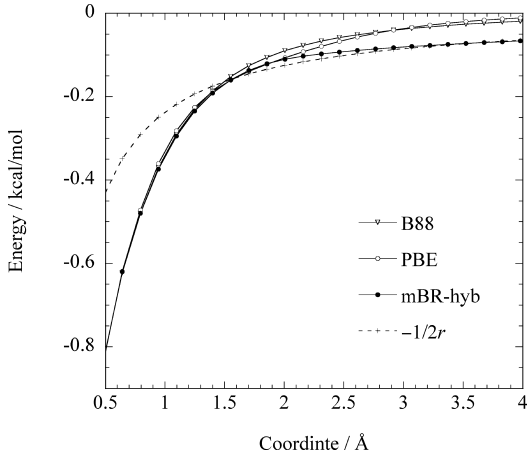


Figure 4. Plots for the exchange potentials given by B88, PBE, and mBR-hyb functionals. The definition of the horizontal axis is the same as in Figure 2. The same electron density used in the construction of Figure 2 is employed for these calculations.

asymptotically behaves as $-1/r$, while that for PBE decays rapidly as the reference electron moves away from the molecule. The origin of such a difference between these two GGA functionals is the same as that of the deviation in the enhancement factors discussed in section 4.2. That is, the bifurcation of the exchange energy densities arises from the difference in the constraints imposed on the functionals. In accord with the B88 fucntional, the curve for mBR-hyb also shows correct asymptotic behavior by virtue of the exchange hole based on the mBR model. We again emphasize that the long-range nature of the exchange energy density is naturally incorporated in the mBR model since it mimics the hole in a hydrogenic atom.

In Figure 4, it is shown that the exchange potential v_x for B88 decays rapidly in contrast to the behavior of the exchange energy density in the long range. This asymptotic behavior coincides with that of the PBE functional. Such an undesirable behavior of the exchange potential of the B88 functional was analyzed in ref 44, and it was proved analytically that any exchange functional $E_{x\sigma}$ with the form of

$$E_{X\sigma}[\rho_\sigma] = \int \rho_\sigma^{4/3} f(x_\sigma)\, d\mathbf{r} \tag{32}$$

does not satisfy the $-1/r$ asymptotic relation when $f(x)$ in eq 32 is constructed so that the exchange energy density behaves asymptotically as $-1/r$. Specifically, v_x for the B88 functional decays non-Coulombically as k/r^2 with a negative constant k. In contrast to B88 and PBE, we observe that the potential for the mBR-hyb functional decays Coulombically in the asymptotic region; however, it recovers only half of the $-1/r$. The $-1/2r$ asymptotic behavior of the mBR-based approach is readily understood by consulting eq 23. In the region of the small electron density, it is verified that the exchange potential $v_{x\sigma}$ is dominated by half of the exchange energy density $U_{x\sigma}$ which behaves as $-1/r$. Thus, it has been demonstrated that the exchange potential of the functional based on the mBR exhibits Coulombic asymptotic behavior of $-1/2r$.

Table 1. Mean Absolute Deviations of Atomization Energies for 35 Molecules in the G2 Set Evaluated by LDA, mBR, mBR-PBE, mBR-hyb, and PBE Exchange Functionals[a]

	LDA	mBR	mBR-PBE	mBR-hyb	PBE
mean abs. deviation	38.5	32.2	16.4	7.9	7.7

[a] Deviations are in the units of kcal mol^{-1}. The LDA exchange is combined with the local correlation term given by Perdew and Zunger.[12] The rest of the functionals are used along with the LYP functional.[39] In the mBR-hyb calculation, the parameters p and c in eqs 19 and 29 are those not optimized (p = 2/3 (0.667), c = 1.0). We refer the readers to Supplementary Table 1 in the Supporting Information for the values of individual molecules.

Table 2. Mean Absolute Deviations of Ionization Potentials for 13 Molecules in the G1 Set Evaluated by LDA, mBR-hyb, and PBE Exchange Functionals[a]

	LDA	mBR-hyb	PBE
mean abs. deviation	0.3	0.2	0.2

[a] Deviations are in the units of eV. The correlation functionals used are common to those in Table 1. In the mBR-hyb calculation, the parameters p and c in eqs 19 and 29 are those not optimized (p = 2/3 (0.667), c = 1.0). We refer the readers to Supplementary Table 2 in the Supporting Information for the values of individual molecules.

4.4. Properties of Small Molecules. In this section, we have assessed the performance of the series of the mBR-based exchange functionals by computing atomization energies, ionization potentials, and proton affinities respectively for 35, 13, and 3 molecules supplied in the G1 and G2 molecular sets.[45,46] The method to determine the asymptotic value of the exponent α_0 of the exchange hole through eqs 17 and 18 necessitates the eigenvalue ϵ_{HOMO} of HOMO to be negative. Unfortunately, ϵ_{HOMO} for anionic molecules are positive in most cases, and hence, the performance check for the electron affinities could not been done. In this sense, our approach lacks robustness and needs to be refined in the procedure to determine the exponent α_0. In Tables 1, 2, 4, and 5, we have only presented the statistics for each property.

Table 3. Proton Affinities for Three Molecules in the G1 Set Evaluated by LDA, mBR-hyb, and PBE Exchange Functionals[a]

system	LDA	mBR-hyb	PBE	expt.
NH_{32}	206.5	210.2	207.1	211.2
$H_{32}O$	167.4	170.7	167.3	173.5
C_2H_2	153.6	158.9	154.9	155.7

[a] Energies are in the units of kcal mol^{-1}. The correlation functionals used are common with those in Table 1. In the mBR-hyb calculation, the used parameters p and c in eqs 19 and 29 are those not optimized (p = 2/3 (0.667), c = 1.0). The experimental values are taken from ref 45, from which the ZPEs are removed. ZPEs are obtained by the same manner as in Table 1.

Table 4. Atomization Energies for 35 Molecules in the G2 Set Evaluated by mBR-hyb, the Original BR Functional, and That Hybridized with the Hartree−Fock (HF) Exchange[a]

	mBR-hyb	BR	BR+HF
mean abs. deviation	4.9	4.3	1.6

[a] Energies are in the units of kcal mol^{-1}. The mBR-hyb exchange functional is used along with LYP correlation functional,[39] while original BR and BR+HF functional are combined with the correlation energy of ref 48. In the mBR-hyb calculation, the parameters p and c in eqs 19 and 29 are those optimized (p = 0.7/3 (0.233), c = 0.15). We refer the readers to Supplementary Table 3 in the Supporting Information for the values of individual molecules.

Table 5. Mean Absolute Deviations of Enthalpies of Formations for 63 Molecules in the G3 Set Evaluated by mBR-hyb and the PBE Exchange Functional[a]

	mBR-hyb	PBE
mean abs. deviation	9.4	18.7

[a] Deviations are in the units of kcal mol^{-1}. The correlation functionals used are common to those in Table 1. In the mBR-hyb calculation, the parameters p and c in eqs 19 and 29 are those optimized (p = 0.7/3 (0.233), c = 0.15). We refer the readers to Supplementary Table 4 in the Supporting Information for the values of individual molecules.

We have supplied "supporting information" for the references to the results for individual molecules.

We have computed the atomization energies by using mBR, mBR-GGA, and mBR-hyb functionals and compared them with experimental values in Table 1. The results by the LDA and PBE functionals have also been provided in the table for comparisons. All exchange functionals except for LDA have been used in combination with the LYP correlation functional.[39] In the LDA calculation, on the other hand, the local correlation functional proposed by Perdew and Zunger (PZ)[12] has been used. Hereafter, we omit the notation of "LYP" or "PZ" specifying the correlation functional for the sake of brevity. Mean absolute deviation (MAD) of the atomization energies from the experimental values has been computed for each functional. As described in the Computational Details, we have checked the error due to the pseudopotentials utilized in the real-space grids scheme by performing all-electron calculations with Gaussian 03 with a sufficiently large basis set (aug-cc-pVQZ). The MAD of the PBE functional derived by Gaussian 03 has been estimated to be 8.2 kcal/mol, which shows good agreement with the value of 7.7 kcal/mol given by our code. It may be reasonable to conclude that the use of the pseudopotential does not seriously affect the energetics at least for these molecules. The MAD for the mBR calculation has been obtained as 32.2 kcal/mol, and it has been found that the mBR approach gives rather better results than the LDA level calculations. However, it is quantitatively far from satisfaction to predict the energetics for the chemical reactions. It has been found that the introduction of the GGA correction to the mBR functional (mBR-GGA) significantly improves the computational accuracies in the atomization energies. The value of MAD has been reduced to 16.4 kcal/mol. It should be noted that the MAD given by mBR-GGA lies in the middle of those given by two sorts of calculations based on the *original* BR approach conducted by Neumann et al.[21] To be specific, they performed calculations by setting the values of γ in eq 4 at $\sim$1 and 0.8. Then, the MAD values were given as 10.3 and 21.1 kcal/mol for the calculations of γ = 1.0 and 0.8, respectively. As described in section 2.4, the relation of γ = 1.0 holds in the exact expression of the spherically averaged Taylor expansion of the exchange hole, while the choice of γ = 0.8 is intended to recover the exchange energy at the homogeneous electron gas limit. We conclude that the mBR-GGA level funtional is almost comparable to the original BR approach in the computational accuracy for the atomization energy. However, it is obvious that the mBR-GGA functional is still less accurate than the PBE functional. We can see in the table that further improvement can be achieved by combining the mBR-GGA with the LDA-based functional (mBR-hyb). The hybridization with the LDA functional reduces the MAD value from 16.4 to 7.9 kcal/mol. Thus, the realization of exchange energy for the uniform electron gas limit at low x_σ is crucial to ensuring the computational accuracy of the functional as suggested in ref 16. Thus, it has been revealed that mBR-hyb is comparable to the PBE functional.

By utilizing the mBR-hyb functional we have also computed the ionization potentials and proton affinities for some molecular systems, the results of which are, respectively, presented in Tables 2 and 3. And comparisons have been made with the results given by LDA and PBE. We can see that the MAD for mBR-hyb is comparable to that for PBE in the calculations of the ionization potentials. As for the proton affinity, the mBR-hyb functional shows slightly better results than PBE, though the number of the samples is very small.

Here, we emphasize that there is still room for further refinement in the mBR-hyb functional. The adjustable parameter p in eq 19 has so far been taken as 2/3, which is simply the maximum of the allowed value which ensures the existence of the real value of r defined by eq 20. In addition, the mixing parameter c in eq 29 has been chosen somewhat arbitrarily as 1.0 to reproduce the exact electronic energy of hydrogen. To optimize these parameters, we have extended the benchmark molecules to the G3 set.[47] Explicitly, 63 molecules in the G3 set have newly been added to the above 35 molecules (98 molecules in total) for the two-parameter fit to the experimental atomization energies and the enthalpies of formation. Molecules in the G3 set that

include the S atom have been excluded from the benchmark test because the pseudopotential for the S atom is not available in the present version of our program. To be specific for molecules in the G3 set that include relatively large hydrocarbons, we have prepared 90 grid points along each direction of the real-space cell to ensure that the wave functions will be enclosed within the cell. This optimization has led to a set of values $p = 0.7/3$ (0.233) and $c = 0.15$. The atomization energies for the above 35 molecules computed by using the parameters have been presented in Table 4 in comparison with the experimental values. It has been demonstrated that MAD has been successfully decreased to 4.9 from 7.9 kcal/mol for the 35 molecules in the G2 set. In the second and third columns in Table 4, we have also presented the results[24] obtained by Becke, who utilized the original Becke−Roussel approach. It should be noted, however, that the computations were performed using the orbitals and densities given at the outset by LDA calculations. The second column shows the results by the BR exchange functional in an unaltered form ($\gamma = 1.0$ in eq 4) in combination with the correlation energy based on the inhomogeneous electron gas model,[48] which gives the MAD from experiment as 4.3 kcal/mol. Thus, it has been found that the parameter-optimized mBR-hyb functional is almost comparable in accuracy to the original BR functional. The data in the third column were obtained by replacing the small fraction of the exchange term by the exact (Hartree−Fock) exchange with mixing parameter $c_X = 0.154$. By mixing the exact exchange in the functional, the MAD value greatly decreased to 1.6 kcal/mol. As described in section 2.2, the size consistency is violated in the present method. We have checked the energy deviation due to the size inconsistency by dissociating the O−H bond in a H_2O molecule. The absolute energy of the sufficiently separated OH and H complex subtracted by the sum of the energies of the constituent fragments has been evaluated as 1.7 kcal/mol. The O−H bond energy has been obtained as 125.0 kcal/mol, and hence, the size inconsistency is not so serious in this case. However, we note that care must be taken for this shortcoming in the functional.

As for the molecules in G3 set, we have computed the enthalpies of formation with the procedure described in the G2 and G3 papers. The enthalpy of formation $\Delta H_f'(A_xB_y)$ at 298 K for a compound such as A_xB_y can be expressed as

$$\Delta H'_f(A_xB_y) = \Delta H_f(A_xB_y) + \{H'(A_xB_y) - H(A_xB_y)\} - x\{H'(A) - H(A)\}_{st} - y\{H'(B) - H(B)\}_{st} \quad (33)$$

where ΔH_f denotes the enthalpy of formation at 0 K, and H' and H stand for the enthalpies at 298 and 0 K, respectively. The corrections for enthalpies of elements are for the standard states of elements and denoted by "st" in eq 33. $\Delta H_f(A_xB_y)$ in eq 33 can be, further, decomposed into

$$\Delta H_f(A_xB_y) = H(A_xB_y) - xH(A) - yH(B) + x\Delta H_f(A) + y\Delta H_f(B). \quad (34)$$

The first three terms in the right-hand side of eq 34 constitute the minus atomization energy corrected by zero-point vibrational energy (ZPE). In the calculation of eqs 33 and 34, the experimental values have been adopted by the elemental enthalpies of formation and by their corrections.[49] Furthermore, it has been assumed that the corrections for the enthalpies of compounds as well as ZPE can be estimated with substantial accuracies by the G3 theory.[50] Then, the errors in the theoretical enthalpies of formation can be reasonably ascribed to those in the atomization energies. In the first column in Table 5, we have presented the enthalpies of formation computed by the mBR-hyb functional with the optimized parameters $(p, c) = (0.233, 0.15)$. The MAD for the 63 molecules in the G3 set has been obtained as 9.4 kcal/mol, where the *n*-octane (C_8H_{18}) molecule has given the maximum absolute error of 23.3 kcal/mol. It should be noted that absolute errors in enthalpies of formation increase with molecular size because of the increase in the atomization energies. For instance, the atomization energy of *n*-octane has been evaluated as 2451 kcal/mol. The third column of Table 5 shows the results given by the Gaussian package using the PBELYP functional with the aug-cc-pVQZ basis set, where the MAD value has been obtained as 18.7 kcal/mol and the maximum absolute error has been obtained as 40.6 kcal/mol for pyrimidine ($C_4H_4N_2$). Furthermore, we have also evaluated the enthalpies of formation by using PBE exchange combined with the PBE correlation functional using the same basis set, which have led to a MAD value of 22.0 kcal/mol. Thus, it has been demonstrated that the mBR-hyb functional with the optimized parameters is superior to a conventional GGA functional for the evaluation of atomization energies of this molecular set.

Last, we present in Figure 5 the results for the computation of the hydrogen-bond (HB) energy curve of a water dimer as a function of the distance between oxygen atoms. The illustration for the geometry of the water dimer has also been drawn in Figure 5. The geometrical parameters of the constituent water molecules have been fixed at those of the TIP4P model ($r_{OH} = 0.9572$ Å, HOH = 104.52°)[51] for the variation of the O−O distance. To construct a reliable standard for the HB energy curve, we have performed a CCSD(T)[52]/aug-cc-pVTZ calculation with the counterpoise corrections by utilizing the Gaussian 03 package. To make comparisons, the computations by the LDA and the PBE functionals have also been carried out by Gaussian 03 in the same manner as the CCSD(T) calculation. In the mBR-hyb calculations, we have employed both parameters that were optimized ($p = 0.7/3$ (0.233), $c = 0.15$) and not optimized ($p = 2/3$ (0.667), $c = 1.0$). In the real-space cell, 90 grids have been placed along each axis with interval $h = 0.118$ Å, and the width of the dense grid has been set at $h/10$. In accord with the findings of a previous work,[53] we observe in Figure 5 that the LDA functional seriously overestimates the binding energy and underestimates the optimum O−O distance. However, the behavior of the HB energy curve is dramatically improved by the introduction of the GGA correction, as exhibited in the PBE curve, which shows comparable behavior with the CCSD(T) result. And, the curve for mBR-hyb with unoptimized parameters lies between those for CCSD(T) and PBE. Importantly, the mBR-hyb curve also shows good agreement with CCSD(T) after the parameter optimizations. Thus, it is demonstrated that

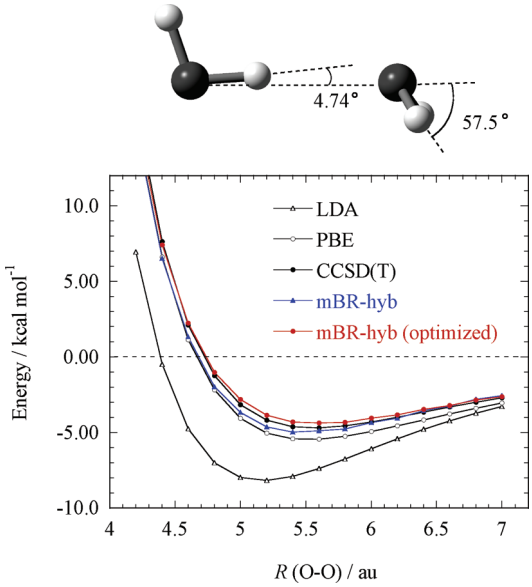


Figure 5. Potential energy curves of a hydrogen bond of a water dimer computed by CCSD(T), LDA, PBE, and mBR-hyb calculations. In the mBR-hyb calculations, both optimized (p = 0.7/3 (0.233), c = 0.15) and nonoptimized parameters (p = 2/3 (0.667), c = 1.0) are used. The horizontal axis represents the O−O distance. The notation of LDA stands for the calculation of Dirac's exchange in combination with the correlation functional proposed by Perdew and Zunger. In the PBE and the mBR-hyb calculations, the correlation energies are estimated by the LYP functional.

the mBR-hyb functional also gives a satisfying result in the computation of the HB energy.

From these test calculations so far, we conclude that the mBR-hyb functional offers energetics with comparable or yet superior accuracy as compared to a sophisticated GGA functional based on the LDA. This greatly encourages further improvements on the mBR-based functionals for practical applications to atoms or molecules.

4.5. Exchange-Hole Function. In this section, we have plotted the exchange-hole distribution of the present approach in comparison with that of LDA and the exact one. We first make a formulation for the exact exchange hole. The exact exchange hole $\rho_{X\sigma}(\mathbf{r}, \mathbf{r}')$ is defined by

$$\rho_{X\sigma}(\mathbf{r}, \mathbf{r}') = \frac{|\rho_{1\sigma}(\mathbf{r}, \mathbf{r}')|^2}{\rho_\sigma(\mathbf{r})} \tag{35}$$

where $\rho_{1\sigma}(\mathbf{r}, \mathbf{r}')$ is the one-body density matrix for spin σ and is simply expressed in terms of the one-electron wave functions; thus,

$$\rho_{1\sigma}(\mathbf{r}, \mathbf{r}') = \sum_i^{\text{occ}} \phi_{i\sigma}(\mathbf{r})\, \phi_{i\sigma}^*(\mathbf{r}') \tag{36}$$

The spherically averaged exchange-hole function $\rho_{X\sigma}^{SA}(\mathbf{r},s)$ can be written as

$$\rho_{X\sigma}^{SA}(\mathbf{r}, s) = \frac{1}{4\pi s^2} \int_{\Omega_s} \rho_{X\sigma}(\mathbf{r}, \mathbf{r} + \mathbf{s})\, d\mathbf{s} \tag{37}$$

where Ω_s denotes the integration over a sphere of radius s centered at a reference point $\mathbf{r}$. On the other hand, the exchange hole based on the mBR model is represented by eq 12, and its spherical average $\rho_{X\sigma}^{SA\text{-}mBR}(\alpha,R;s)$ is obtained by simple algebraic manipulations; thus,

$$\rho_{X\sigma}^{SA\text{-}mBR}(\alpha, R;s) = \frac{\alpha^{1/2}}{4\pi^{3/2}Rs}(\exp\{-\alpha(R - s)^2\} - \exp\{-\alpha(R + s)^2\}) \tag{38}$$

It is useful to note that eq 38 corresponds to eq 17 of ref 18 that utilizes the Slater-type function as an exchange-hole model. We also note that (α, R) in eq 38 has been determined from the spin density $\rho_\sigma(\mathbf{r})$ as described in section 2.2 where the scaling parameter p in eq 19 has been set at 0.233. In the LDA approach, the exchange hole is spherically symmetric around $\mathbf{r}$ and is given by

$$\rho_{X\sigma}^{SA\text{-}LDA}(\mathbf{r}, s) = 9\rho_\sigma(\mathbf{r})\left(\frac{\sin t - t\cos t}{t^3}\right)^2 \tag{39}$$

where

$$t = (6\pi^2\rho_\sigma(\mathbf{r}))^{1/3}s \tag{40}$$

We have plotted the exchange-hole distributions expressed by eqs 38, 39, and 40 for various positions of the reference electron in a molecule to make comparisons. Furthermore, we have also performed calculations with the mBR functional, where the parameters (α, R) in eq 38 have been determined by the procedure proposed in the original Becke−Roussel approach. Explicitly, (α, R) are obtained by imposing the conditions that zeroth- and second-order terms in the Taylor expansion of eq 4 coincide with those of the exchange-hole model defined by eq 12. The computational procedure for the Gaussian-type function is parallel to that for the Slater-type one. In the following, we term this approach the original mBR.

For the construction of the exchange hole, we have employed a water molecule with the same geometry that was used in section 4.2. In the real-space grid approach, 90 grids have been placed along each axis with grid spacing h = 0.0679 Å, and the width of the dense grids has been set at h/5. We have constructed the exchange-hole functions by using the same spin density and one-electron wave functions obtained from the outset by the B88 exchange functional without electron correlations. In Figure 6a−c, we have compared the behaviors of exchange-hole functions for three different positions of the reference electron. In Figure 6a, the reference point has been placed at the oxygen atom. It appears in the contour map that the exchange-hole function is dominated by the $2a_1$ orbital of the water molecule. The exchange-hole functions for mBR and LDA show good agreement with the exact one, while the original mBR model slightly deviates from the others. Anyway, it has been demonstrated that these three approaches can properly simulate the behavior of the exact exchange hole. In Figure 6b, the reference point has been placed below the oxygen atom along the symmetry axis of the molecule. In this case, we recognize that the exchange hole distribution is character-

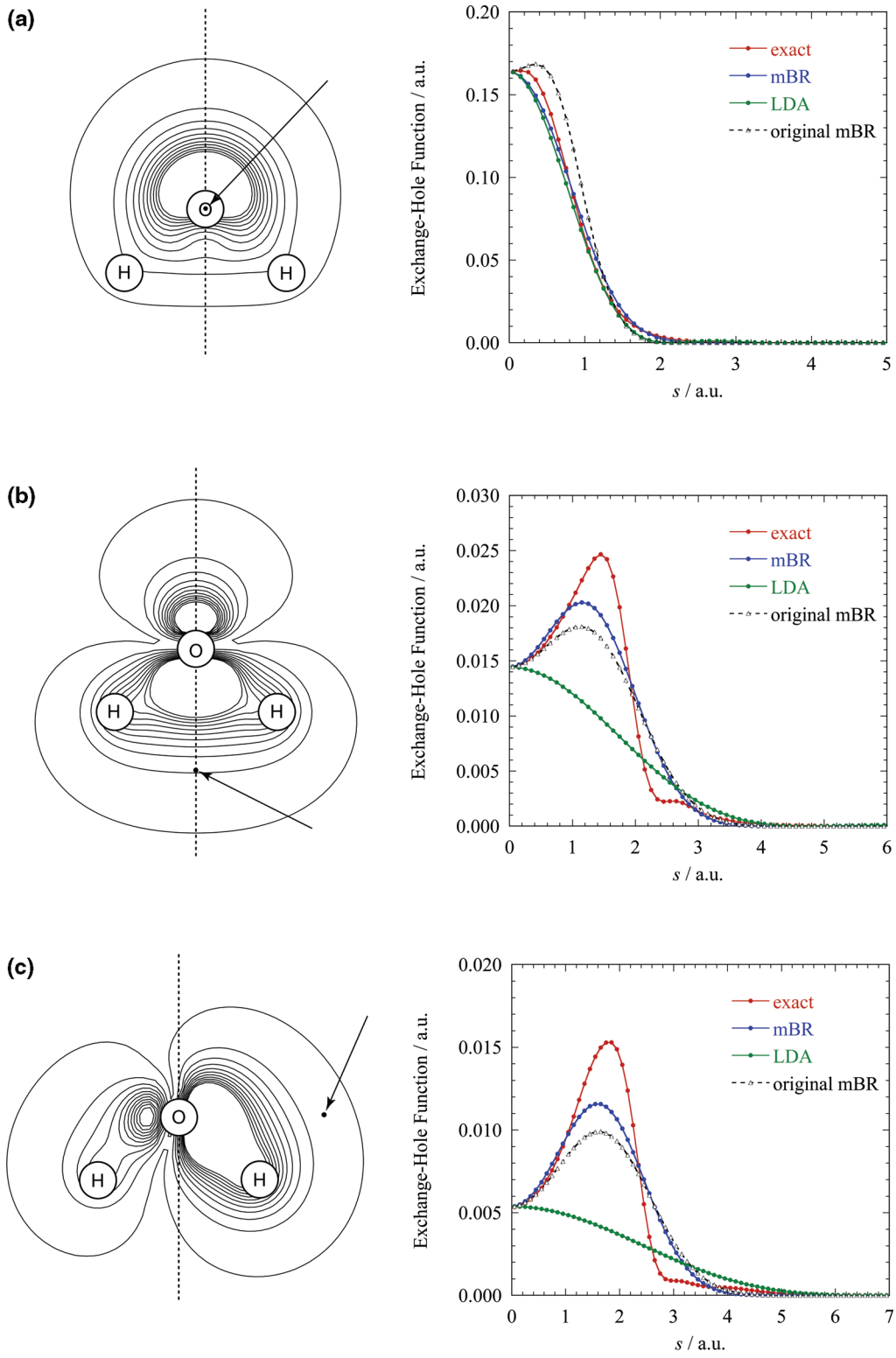


Figure 6. Contour plots of the exchange-hole functions for the reference electron placed at various positions on the molecular plane of H_2O (left). Spherically averaged exchange holes as functions of the distance *s* from the reference point (right). Arrows in the contour map indicate the positions of the reference electrons. The value of the outermost contour line is 0.001 au, and the interval is set at 0.01 au.

ized by the $3a_1$ orbital. It is shown in the plot that the LDA approach fails to reproduce even the qualitative behavior of the exact exchange hole. This originates from the fact that the maximum peak of the exchange hole always locates at the reference point in the homogeneous electron gas. On the other hand, we observe that the mBR approach successfully simulates the behavior of the exact exchange hole, though the detailed structures of the hole cannot be reproduced. This directly demonstrates the advantage of the mBR approach that it allows the exchange hole being attached at the molecule even when the reference electron is moved far apart from the system. We have also plotted the exchange

hole function given by the original Becke-Roussel procedure. The peak position as well as the overall behavior of the original mBR shows good agreement with those of the mBR exchange hole. In Figure 6c, the position of the reference electron has been chosen so that the exchange hole is characterized by the $1b_2$ orbital. Again, we see that the mBR approach can illustrate the appearance of the exact exchange-hole distribution in contrast to LDA. The mBR approach also shows excellent agreement with the original mBR.

5. Summary and Conclusions

The point of the BR approach is to mimic the exchange hole function at a given reference point by an electron distribution of a hydrogenic atom. This allows the exchange hole to be attached to the molecule even when the reference electron is placed far apart from the system, in contrast to the LDA approach, which naturally improves the asymptotic behavior of the exchange energy density. In this paper, we have proposed a series of exchange functionals based on the mBR model for the purpose of finding a new route to the exchange functional. Our approach to the simple realization of the mBR-based functional consists of three steps. The first step is to determine the parameters (α, R) that specify the distribution of the mBR exchange hole with respect to the reference point. The width α of the hole is to be obtained from its asymptotic values α_1 and α_0 through eq 19. Then, the distance r between the reference point and the exchange hole is readily computed from the constraint that the depth of the hole is equal to the spin density at the reference point. In the second step, we make a GGA correction to the functional obtained in the first step (mBR-GGA). More specifically, as expressed in eq 28, the conventional GGA formula has been adopted to take into account the gradients of the spin density as well as the mBR exchange hole. Third, the gradient-corrected mBR functional thus obtained is combined with the LDA approach as in the form of eq 29 to reproduce the exchange energy at the homogeneous electron gas limit (mBR-hyb).

We have examined the behaviors of the enhancement factors of the mBR-based exchange functionals with respect to the distance between the reference electron and a molecular system. It has been demonstrated that the mBR-hyb functional shows an excellent overall behavior in accord with B88. The exchange energy densities of the mBR-based approaches have shown correct asymptotic behaviors of $-1/r$ by virtue of the fact the electron density of a hydrogenic atom is taken as a model of the exchange hole. Further, the exchange potential in mBR-hyb has also been shown to decay Coulombically in contrast to B88 and PBE; however, it recovers only half of the $-1/r$ potential in the asymptotic region. We have assessed the performance of the series of the mBR exchange functionals combined with the LYP correlation functional by computing several properties of the small molecules in the G1, G2, and G3 sets. It has been found that the computational accuracy for atomization energies can be systematically improved by the three steps noted above. It has been demonstrated that the mBR-hyb functional is almost comparable in accuracy to the GGA functional of PBE. Optimization of only two adjustable parameters in mBR-hyb functional has shown to provide much better results in the atomization energies and enthalpies of formation. The MAD value for the atomization energies of 35 molecules in the G2 set has been evaluated as 4.9 kcal/mol by the mBR-hyb functional with optimized parameters. And the MAD for the enthalpies of formations of 65 molecules in the G3 set has been computed as 9.4 kcal/mol. The calculations for the water dimer have revealed that the mBR-hyb is adequate enough to reproduce rather weak interactions such as hydrogen bonds.

The results of the test calculations obtained so far is very encouraging for further improvement of the mBR-based functional. In spite of the success, the present method clearly has deficiencies in some respects. First of all, the procedure to determine the exponent α_0, that utilizes the energy level of the HOMO, must be refined for the applications to anionic systems. The use of the orbital energy also leads to an undesirable property in the functional, that is, the lack of size consistency. Second, the parameters p and c which appear, respectively, in eqs 19 and 29 should be optimized for molecular systems involving all electrons to avoid the influence of the pseudopotentials used in the present calculations. We conclude that the route to develop the exchange energy functional that begins from the BR model is worth consideration as a potential candidate for the establishment of the exchange functional suitable for the applications to atoms or molecules.

Acknowledgment. This work is supported by a Grant-in-Aid for Scientific Research on Priority Areas (Nos. 18031022 and 21118512) from the Ministry of Education, Science, Sports and Culture of Japan. H.T. would like to thank K. Kubota and K. Maruyama for their help in preparing the data in Tables 1∼5. H.T. is also grateful to Prof. N. Matubayasi in Kyoto university for his continuous encouragement of this work.

Supporting Information Available: Additional tables are provided. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dirac, P. A. M. *Proc. Cambridge Phil. Soc.* **1930**, *26*, 376–385.

(2) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385–390.

(3) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–1138.

(4) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–871.

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

(6) Koch, W. Holthausen, M. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, Germany, 2001.

(7) Herman, F.; Van Dyke, J. P.; Ortenburger, I. B. *Phys. Rev. Lett.* **1969**, *22*, 807–811.

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

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

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

(11) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401–146405.

(12) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.

(13) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *124*, 091102.

(14) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(15) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(16) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2006**, *123*, 062201.

(17) Becke, A. D. *Int. J. Quantum Chem.* **1983**>, *23*, 1915–1922.

(18) Becke, A. D.; Roussel, M. R. *Phys. Rev. A* **1989**, *39*, 3761–3767.

(19) Becke, A. D. *J. Chem. Phys.* **2003**, *119*, 2972–2977.

(20) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.

(21) Neumann, R.; Nobes, R. H.; Handy, N. C. *Mol. Phys.* **1996**, *87*, 1–36.

(22) Neumann, R.; Handy, N. C. *Chem. Phys. Lett.* **1995**, *246*, 381–386.

(23) Beck, T. L. *Rev. Mod. Phys.* **2000**, *72*, 1041–1080.

(24) Becke, A. D. *Int. J. Quantum Chem. Quantum Chem. Symp.* **1994**, *28*, 625–632.

(25) Bahmann, H.; Ernzerhof, M. *J. Chem. Phys.* **2008**, *128*, 234104.

(26) Katriel, J.; Davidson, E. R. *Proc. Natl. Acad. Sci.* **1980**, *77*, 4403–4406.

(27) Almbladh, C.-O.; von Barth, U. *Phys. Rev. B* **1985**, *31*, 3231–3244.

(28) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.

(29) Szabo A. Ostlund, N. S. *Modern Quantum Chemistry, Introduction to Advanced Electronic Structure Theory*; Macmillan: New York, 1982.

(30) Takahashi, H.; Hori, T.; Wakabayashi, T.; Nitta, T. *Chem. Lett.* **2000**, *3*, 222–223.

(31) Takahashi, H.; Hori, T.; Wakabayashi, T.; Nitta, T. *J. Phys. Chem A* **2001**, *105*, 4351–4358.

(32) Hori, T.; Takahashi, H.; Nitta, T. *J. Theoret. Comput. Chem.* **2005**, *4*, 867–882.

(33) Chelikowsky, J. R.; Troullier, N.; Saad, Y. *Phys. Rev. Lett.* **1994**, *72*, 1240–1243.

(34) Chelikowsky, J. R.; Troullier, N.; Wu, K.; Saad, Y. *Phys. Rev. B* **1994**, *50*, 11355–11364.

(35) Barnett, R. N.; Landman, U. *Phys. Rev. B* **1993**, *48*, 2081–2097.

(36) Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425–1428.

(37) Ono, T.; Hirose, K. *Phys. Rev. Lett.* **1999**, *82*, 5016–5019.

(38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.05; Gaussian, Inc.: Pittsburgh, PA, 2003.

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

(40) DeFrees, D. J.; Levi, B. A.; Pollack, S. K.; Hehre, W. J.; Binkley, J. S.; Pople, J. A. *J. Am. Chem. Soc.* **1979**, *101*, 4085–4089.

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

(42) Lieb, E. H.; Oxford, S. *Int. J. Quantum Chem.* **1981**, *19*, 427–439.

(43) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890–890.

(44) van Leeuwen, R.; Baerends, E. J. *Phys. Rev. A* **1994**, *49*, 2421–2431.

(45) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622–5629.

(46) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

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

(48) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 1053–1061.

(49) Chase, M. W., Jr.; Davies, C. A.; Downey, J. R.; Frurip, D. J.; McDonald, R. A.; Syverud, A. N. *J. Phys. Chem. Ref. Data* **1985**, *14*, Suppl. No. 1.

(50) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764–7776.

(51) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(52) Purvis, G. D., III; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

(53) Laasonen, K.; Csajka, F.; Parrinello, M. *Chem. Phys. Lett.* **1992**, *194*, 172–174.

CT900416X

JCTC Journal of Chemical Theory and Computation

Study of the Ground State Dissociation of Diatomic Molecular Systems Using State-Specific Multireference Perturbation Theory: A Brillouin−Wigner Scheme

Uttam Sinha Mahapatra

Department of Physics, Taki Government College, Taki, North 24 Parganas-743429, India

Sudip Chattopadhyay*

Department of Chemistry, Bengal Engineering and Science University, Shibpur, Howrah 711103, India

Rajat K Chaudhuri

Indian Institute of Astrophysics, Bangalore 560034, India

Received August 26, 2009

Abstract: The size-extensive second-order state-specific (or single root) multireference (MR) perturbation theory (SS-MRPT) in the Brillouin−Wigner (BW) form using Møller−Plesset perturbative evaluations of orders up to 2 [termed as SS-MRMPPT(BW)] presents a viable, as well as promising, approach to include both nondynamic and dynamic correlations in the study of the bond-stretching (in multireference/quasidegenerate situations) of molecular species with a manageable cost/accuracy ratio. It combines numerical stability in the presence of an intruder state problem with strict size consistency (when localized orbitals are used). In this paper, the SS-MRMPPT(BW) method has been shown to properly break the bonds (in the ground state) of several diatomic molecules (such as F_2, Cl_2 and Br_2, and BH) that have posed a severe challenge to any many-body theoretical approach due to the presence of quasidegeneracy of varying degrees in the ground state. A comparison of the resulting potentials with the various theoretical results reveals that the method represents a valuable tool that is capable of properly accounting even for very strong quasidegeneracies, while also performing well in nondegenerate situations. In this work, we have also calculated spectroscopic constants (such as equilibrium bond lengths, vibrational frequencies, and dissociation energies) of the ground state of these molecular systems. The SS-MRMPPT spectroscopic constants are compared with the most accurate available *ab initio* calculations and other theoretical estimates of previous works to calibrate the efficacy of the method. For the sake of completeness, we also compare the computed spectroscopic constants with the experimental observations. The accuracy of computed spectroscopic parameters appears to be rather consistent over a multitude of systems for various basis sets. The SS-MRMPPT enables quantitatively accurate and computationally affordable analysis of potential energy surfaces and spectroscopic constants of various multireference systems in the ground state. It is particularly visible for spectroscopic parameters and nonparallelism error (NPE) calculations. The calculations further reveal that the SS-MRMPPT(BW) method compared to its Rayleigh−Schrödinger counterpart [SS-MRMPPT(RS)] provides a more accurate and consistent solution for the whole dissociation path and spectroscopic constants.

I. Introduction

Accurate quantum mechanical calculations of molecular electronic energy variations on potential energy surfaces (PESs) are essential in many contexts [such as equilibrium geometries, transition states, force constants, etc.]. The study of molecular electronic energies along bond breaking−making paths also provides key information needed in reaction kinetics and various spectroscopic processes. Thus, it is worthwhile to try and calculate the PES at a level of accuracy that, hopefully, allows a comparison with experimental data (spectroscopic quantities).

The computation of a smooth PES requires a balanced treatment of dynamic and static correlation effects, and thus a genuine multireference (MR) formalism is warranted. Among the possible types of MR technologies available to compute the PES, we consider here MR-based perturbation theory (MRPT),[1−17] which is, in general, computationally cost-effective in describing dynamical correlation in the presence of electronic degeneracy. The conventional effective

* To whom correspondence should be addressed. E-mail: sudip_chattopadhyay@rediffmail.com.

10.1021/ct900452n
Published on Web 02/12/2010

Hamiltonian MR approaches based on complete model space (CMS) suffer convergence difficulties once intruder states intervene while studying PES, which spoil the quality of the spectroscopic constants. To overcome this bottleneck, a state-specific multireference (SSMR) approach—targeting only the state of interest unencumbered by intruders—is an inherently useful strategy. Over the past few decades, many promising methods have emerged [for an overview, see ref 18]. Among the various intruder-free MRPT methods, multireference Møller−Plesset perturbation theory (MRMPPT),[3] complete active space perturbation theory (CASPT),[5,6] and second-order *n*-electron valence perturbation (NEVPT2) theory[11] approaches have been established as very efficient methods for computing the PES of any state (regardless of charge, spin, or symmetry) with satisfactory and consistent accuracy [see also ref 19]. The convergence of the MRPT with MP partitioning is not free from objections.[20−23] It must be admitted that CASPT2 can handle large active space leading to a satisfactory accuracy. However, it has two serious limitations which may invite inconsistent behavior of the method. One is a lack of strict size extensivity and the other, more importantly, is that it is occasionally subjected to the intruder state problem, causing divergences in the perturbation series.[24,25] MRMPPT[3] and MCQDPT[4] are also not rigorously size-extensive.[24,25] The main features of recently developed NEVPT2[11] theory are size consistency and the absence of intruder states. In our view, a state specific (or single root) MR method that is size-extensive as well as size-consistent is essential and crucial to getting correct results of dissociation PESs and associated spectroscopic parameters [see ref 26 regarding this aspect].

We also mention other developments in the context of MRPT methods which are able to deal with the PES. There have been significant contributions by Wolinski and Pulay,[27] Murphy and Messmer,[28] Dyall,[29] and most recently by Robinson and McDouall.[30] In this context, we also mention the development of multiconfiguration perturbation theory by Surján and his group.[31] Surján and Rosta[32] have investigated the MRPT using the APSG (antisymmetrized product of strongly orthogonal geminal) reference state. To increase the computational effectiveness, a number of groups[30,33−35] have investigated the possibility of avoiding the CASSCF step, by using orbitals obtained from simpler methods to define the active spaces for use in multireference perturbative treatments. The earlier CIPSI (Configuration Interaction with Perturbation Selection Iteratively) method[36] may be viewed as a forerunner of many of the more recent MR perturbation theories. This method can be viewed as a second order perturbation correction to CI energies via diagrammatic techniques using multiconfigurational zeroth order wave functions.

In recent times, Mukherjee and co-workers[14,15,18] have introduced a rather elegant size-extensive and size-consistent MRPT approach to tackle the intruder state problem [referred to as state specific second-order MRPT, SS-MRPT]. In SS-MRPT formulation, they used the Jeziorski−Monkhorst (JM) ansatz[37] in a state specific fashion. The SS-MRPT utilizes a multipartitioning strategy.[10,38] The SS-MRPT is quite rich in its structure in the sense that it can be viewed as versions of both a Rayleigh−Schrödinger (RS) and a Brillouin−Wigner (BW) perturbation theory, depending on the expansion strategy [we use the nomenclature SS-MRPT(RS) and SS-MRPT(BW), respectively]. The SS-MRPT method has *all the attractive features* of the parent SS-MRCC method.[39] The SS-MRPT method obeys satisfactorily the logical and practical requirements of a good MRPT approach such as the following: (i) it is rigorously size-extensive and size-consistent (when localized orbitals are used); (ii) it is generally applicable to a wide class of problems within one framework, i.e., not dependent on specific choices of configurations; (iii) it bypasses the *intruder* state problem; (iv) it is efficient and cost-effective; (v) the model can properly treat dissociation of fragmentation problems (reactions) in a spin-pure way for closed as well as open shells (spin symmetry is essential to the proper description of bond breaking); (vi) there is flexibility of using the coefficients of reference functions in either a relaxed or an unrelaxed mode under the effect of the perturbation; (vii) it is able to calculate energies of similar quality in a wide domain of geometries. As the SS-MRPT method is designed to facilitate the relaxation of the reference function, it thus deals with the mixed-states problem characterized by large changes in the relative contributions of the coefficients of reference functions in an exact (correlated final) wave function compared to the zero-order function. The SS-MRPT approach is able to model any region of the PES of a molecular system (with closed-/open-shell and singlets/nonsinglets model functions) even when the traditional effective Hamiltonian based MR methods fail due to intruders.[14,15,18,40] More recently, a number of researchers have adapted the SS-MRPT approach to a production-level implementation;[41] hitherto the derivation has been formulated in terms of spin−orbitals. One of the main objections of the spin−orbital based formulation exists in the incorrect treatment of spin eigenstates. As a word of caution, it is important to note that the SS-MRPT method is useful as long as the target state energy is well separated from the virtual functions. This is generally true for the ground and low-lying states.[40] Our application in this paper is based on the problem of describing PESs of molecular systems in their ground states.

The main difficulty in CAS-based MR calculations arises from the use of CAS in constructing many-electron wave functions due the exponential increase of the size of the model space when one increases the number of active orbitals. The main relevant criticism of the theories based on the wave operator of the Jeziorski−Monkhorst type is its prohibitively increasing number of amplitudes as a function of the size of the model space, since the cluster operator is defined with respect to each reference determinant. The computational scaling, however, is not favorable if the number of the reference determinants is large. Beyond that, for each cluster operation (specific for the target state), one must solve a set of coupled equations which scales with the CAS size. In contrast, the ansatz for internally contracted methods is more compact. Despite the computational benefit of the internal contraction, internally contracted state-specific multireference formalisms may suffer due to internal contraction of the wave function in the reference space. In the

case of computation of the interaction between close lying states, or of weak crossings, some of the spurious effects induced by the internal contraction[42] may affect the stability and accuracy of this type of method. It is worth mentioning that CASPT2[5,6] can be implemented in terms of either internally contracted configurations, configuration state functions, or some combination of the two, whereas MRMPPT[3] is formulated in terms of configuration state functions. In this context, we can also mention the work on the reduced model space method in MRPT of Staroverov and Davidson.[43]

In order to corroborate the SS-MRPT with some applications, we have chosen to start with the implementation of its various variants. Instead of following an RS expansion based MP-type philosophy in SS-MRPT [named SS-MRMPPT(RS)], one may also use the BW-based MP-type [termed SS-MRMPPT(BW)] spirit while retaining the main advantages of size extensivity and size consistency rigorously in an intruder free way. Our multipartitioning scheme for RS and BW follows the ideas advanced by Mahapatra et al.[14] and Ghosh et al.,[15] respectively. There have been several numerical studies illustrating the ability and power of the SS-MRMPPT(RS) method to describe single, double, and triple bond breaking on singlet and nonsinglet PESs,[14,40] but little is known about the performance of the SS-MRMPPT(BW) approach in calculations of bond breaking in molecular species.[14,15] The pre-eminent success of the SS-MRMPPT(RS) method in treating electron correlation for many types of systems prompted us to apply the SS-MRMPPT(BW) scheme. The main goal of the present investigation and of the analysis is to examine the performance of the SS-MRMPPT(BW) method in the difficult quasi-degenerate situation arising from the bond stretching and to compare our results with other approaches, particularly those belonging to the general category of the MRMP method (such as CASPT2, MRMPPT, BWPT, etc). It would be very constructive if we are able to incorporate a comparative study of the results of the SS-MRMPPT method with respect to the other nonperturbative methods in each case. In this work, we also inspect explicitly the convergence issue with respect to the correlation treatment and the basis set for the SS-MRMPPT(BW) method along with its RS version. However, the application of the SS-MRMPPT(RS) is computationally less demanding than SS-MRMPPT(BW) at the cost of accuracy.

In this paper, we investigate the ground state of the X_2 [X = F, Cl, and Br] and BH molecules as our benchmark systems, as they are the prototype systems for a variety of spectroscopic and reaction dynamics studies. The importance of the X_2 systems as laser media, as well as their unusual behavior, provides motivation for trying to understand them better. The molecular dihalogens exhibit interesting yet only partially understood molecular properties, one being the apparent scrambling of the order in bond dissociation energies. As one descends along group VIIA of the periodic table, there should be an expected decrease from fluorine to iodine; curiously, the chorine molecule has the highest bond dissociation energy. The traditional explanation lies in positing the fact that the fluorine atom has a very small size, with high electron density over it, resulting in molecular fluorine being a stretched molecule (with a more than expected bond distance, as is envisaged in fact) owing to the high lone-pair−lone-pair repulsion. The fluorine molecule shows different electronic behavior in relation to chlorine and bromine molecules. It is already known that F_2 is unbound at the Hartree−Fock (HF) level. In contrast to fluorine, chlorine and bromine molecules are bound at the HF level. The situation dramatically changes for the CASSCF (Complete Active Space−Self Consistent Field) wave function. However, despite the improvement, the calculated bonding energy is still disappointingly small, as CASSCF does not treat the dynamic correlation, which is important for a correct description. Relying on the generalized valence bond method, recently Barbosa and Barcelos[44] have put forth an additional understanding of this phenomenon. It is now well-known that to compute the PES within "chemical accuracy", sextuple excitations with respect to single-determinant reference functions are needed even for the single-bond (F−F) breaking of the F_2 molecule.[45] This is the main reason for the problems faced by the SR-based methods.

Along the ground state dissociating reaction path of X_2 and BH, the starting reference function changes multiconfigurationally and encounters the perennial "intruder state problem", which prompted us to undertake the present study of X_2 and BH dissociation. It is to be remembered that, when the "intruder state" makes a contribution to the target state, removing it from the perturbation expansion in the region of the singularity invites in a noticeable error in the computed perturbative energy in the context of an effective Hamiltonian-based approach(es). We have observed that the SS-MRMPPT formalism does not break down when generating a ground state dissociation energy surface for X_2 and BH. In this work, we also focus on the calculation of the spectroscopic constants through the computation of nonrelativistic PES for the dissociation of the X_2 molecule into two X atoms in the ground electronic state (belongs to the $^1\Sigma_g^+$ representation) via the SS-MRMPPT(BW) [and also SS-MRMPPT(RS)] approach using various basis sets. We consider the same for the ground state of the BH molecule also. At this point, we want to state that the most detailed experimental information regarding energies along the entire reaction paths is available in diatomic molecules by virtue of their spectroscopic constants, and thus they furnish exacting tests for methods attempting to describe PESs. Actually, theoretical computation of vibrational frequencies has become almost "a must" for experimental spectroscopists these days, as it helps to interpret and assign experimental infrared or Raman spectra, especially in difficult and questionable cases. The SS-MRMPPT spectroscopic constants (evaluated by fitting the energy surfaces or curves to cubic polynomials) are found to be in reasonable agreement with the corresponding experimental values. In the present work, we calculate the dissociation energies as the difference in energy at an asymptotically large distance and the fitted energy minimum.

Chlorine and bromine introduce a new level of complexity to our computational methods because they are so large. Chlorine and bromine have 17 and 35 electrons, respectively,

and taking into consideration every electron for these atoms costs a great deal computationally. Also, the core electrons have quite high energy so that there is the possibility of relativistic effects on the core electrons. In systems containing heavy atoms, such as chlorine or bromine, relativistic effects and spin−orbit coupling cannot be neglected (in some cases even in the first-order perturbative treatment) to get correct results. Since the ground state ($^1\Sigma_g+$) has zero angular and spin momentum, it is not subject to spin−orbit splitting; however, relativistic effects play an important role. These relativistic effects are not taken into consideration under standard levels of theory. In our calculations presented in this paper, the inner-shell electron correlation and relativistic effects have been disregarded. In our remaining discussion, we will mainly focus on the analysis of the basis set and a balanced treatment of dynamical as well as nondynamical correlation effects, the key to a successful description of molecular PESs involving bond making and breaking.

The SS-MRMPPT(BW) approach has been shown to work slightly better and be consistent in comparison to the SS-MRMPPT(RS). It is important to note that the convergence of the correlation energy with basis set size of the SS-MRMPPT(BW) is better than the RS counterpart, as it should be. A close observation of the numerical performance of the SS-MRMPPT methods exhibits that the overall performance of SS-MRMPPT(BW) is better and more consistent over its RS counterpart. This paper is not to advocate replacement of the SS-MRMPPT(RS) approach with the SS-MRMPPT(BW) one; rather, it is to throw light on the role of the scheme of perturbative expansion in the treatment of the coupling term maintaining size extensivity. The SS-MRPT method is generalizable to SS-MRENPT in RS and BW expansion. In this context, we remark that the SS-MRENPT approach is endowed with size extensivity and size consistency properties in contrast to the traditional MRPT method with EN partition. In the present work, the SS-MRENPT with RS and BW expansion has been applied to investigate the F_2 system. The results of the SS-MRPT with an EN partition are not too good. The ground state PESs of F_2 via SS-MRMPPT have already been published by Mahapatra et al.[14] with the EN-type partitioning scheme (using RHF orbitals). However, that work was based on relatively small basis sets. We have used sequences of the correlation-consistent cc-pVXZ basis sets (with spherical components) of Dunning in calculations described in this paper. The basis sets used in this paper were taken from ref 46.

Before embarking on the numerical performance of the SS-MRPT method, we first discuss in section II some salient methodological aspects of the SS-MRPT method, which are pertinent for numerical discussion. The numerical results are presented and discussed in section III. A summary and outlook is given in section IV.

II. Brief Résumé of Theory and Discussion

Before presenting the results of the present work, we shall give a brief introduction of the SS-MRPT theory. It is not our purpose here to describe the detailed derivation of the SS-MRPT method starting from the mother theory, the SS-MRCC formalism. Such details can be found elsewhere.[14,18,39] Considerable theoretical and computational progress has been achieved for a companion perturbation theory (SS-MRPT) for state-specific multireference coupled cluster methods of Mukherjee and co-workers.[18,39,41] The formal development of SS-MRPT has been based on the JM ansatz. In the JM ansatz, the wave function has the form

$$|\psi\rangle = \sum_{\mu} \exp(T^{\mu})|\phi_{\mu}\rangle c_{\mu}$$

where the combining coefficients of the model space (MS) functions c_μ's are a priori unknown. The reference function ψ_0 is a combination of the MS function (configuration state functions, CSF) ϕ_μ, called the complete model space (CMS):

$$\psi_0 = \sum_{\mu} c_{\mu}\phi_{\mu} \tag{1}$$

Here, the cluster operator, T^μ, acting on ϕ_μ creates a set of virtual functions, $\{\chi_i^\mu\}$. Every CSF is associated with its own cluster excitation operator to take care of differential correlation and dynamical correlation effects, instead of applying one universal operator to the whole reference function. This makes the approach based on this ansatz the method of choice for computing the state energies *per se*. Inserting the ansatz into the Schrödinger equation yields[37]

$$(H - E)\sum_{\mu} \exp(T^{\mu})|\phi_{\mu}\rangle c_{\mu} = 0$$

If all the parameters are independent, there is redundancy in the ansatz in the sense that some excited CSFs can be reached in multiple ways by the linear excitation of some reference CSFs. Thus, the use of the JM ansatz in a state specific fashion introduces an inherent redundancy problem. Mukherjee and co-workers tackled the obstacle and solved the amplitude equations by using some physically motivated sufficiency conditions.[39] The SSMR approach of Mukherjee and co-workers explicitly contains the eigenvector coefficients in contrast to both state-universal[37] and Brillouin−Wigner based methods.[12] In recent years, BW methods have been applied to the many-electron systems in a state specific formulation. For all practical purposes, actual perturbative computations require the truncation of the perturbative expansion, and this truncation is only meaningful if the perturbation series either converges rapidly or at least converges rapidly in an asymptotic sense. The general problem of spin-adaptation using the JM ansatz based MR methods is quite a nontrivial and a rather involved exercise.[47] The spin-adaptation of the SS-MRPT method has been achieved (i) by invoking suitable spin-free unitary generators to define the cluster operators and (ii) by considering the entire portion of the highest closed-shell component, $\phi_{0\mu}$, of a model function ϕ_μ as the vacuum to define all the excitations on ϕ_μ in normal order [see ref 18 for the spin-free development of SS-MRPT in detail]. In the SS-MRPT development, intermediate normalization is used, making the introduction of a complete active space essential (the amplitudes corresponding to the internal excitations are set to zero). It is also worth mentioning that the assumption of a complete model space enables proving the connectivity of

the cluster operator making the method scale correctly with the number of electrons (size-extensive). This is an important and nontrivial result. It should be noted that, because of the lack of invariance of the SS-MRPT with respect to active orbitals, strict size consistency can be demonstrated only for localized orbitals (which would also be the case for the analogous perturbation theory).

To discuss the structural features of cluster finding equations of the SS-MRPT approach, we recall the following quasi-linearized form of the parent SS-MRCC theory:[39]

$$\langle\chi_l^\mu|H|\phi_\mu\rangle c_\mu + [(\sum_m \langle\chi_l^\mu|H|\chi_m\rangle - \langle\phi_\mu|H|\phi_\mu\rangle\delta_{lm})\langle\chi_m^\mu|T^\mu|\phi_\mu\rangle]c_\mu - [\sum_\nu \langle\chi_l^\mu|T^\mu|\phi_\mu\rangle\varepsilon]c_\mu + [\sum_\nu \langle\chi_l^\mu|T^\nu|\phi_\mu\rangle\tilde{H}_{\mu\nu}]c_\nu = 0 \forall l,\mu \tag{2}$$

Here, $\{\chi_l^\mu\}$ (for all μ) are the set of virtual functions spanning the space complementary to the space of ψ_0, $\tilde{H}_{\mu\nu} = \langle\phi_\mu|H|\phi_\nu\rangle + \langle\phi_\mu|\overline{HT^\mu}|\phi_\nu\rangle$. In the development of SS-MRPT, one can treat ε a dependent from our choice of $\tilde{H}_{\mu\nu}$, depending on the RS or BW mode of formulation, but not on a specific partitioning strategy. For the RS version, one can choose $\tilde{H}_{\mu\nu}$ as $H_{\mu\nu}$, while the second order effective pseudo-operator $\tilde{H}_{\mu\nu}^{(2)}$ can be chosen for the BW scheme. The partitioning of H affects the second and third terms of eq 2. Thus, from the very mode of development of the SS-MRPT method, one can observe that the formalism provides a completely independent treatment of the size-extensivity correction term and partitioning of the total Hamiltonian. Hence, one can consider several schemes for the development of the MRPT-like approach from the parent SS-MRCC theory.

For actual applications, the form of the working equations for the first-order cluster amplitudes can be written as

$$t_\mu^{l(1)} = \frac{H_{l\mu}}{[E_0 - H_{ll}]} + \frac{\sum_\nu^{\nu\neq\mu} \langle\chi_l^\mu|T^{\nu(1)}|\phi_\mu\rangle H_{\mu\nu}(c_\nu^0/c_\mu^0)}{[E_0 - H_{ll}]} \tag{3}$$

for RS, and

$$t_\mu^{l(1)} = \frac{H_{l\mu}}{[E - H_{ll}]} + \frac{\sum_\nu^{\nu\neq\mu} \langle\chi_l^\mu|T^{\nu(1)}|\phi_\mu\rangle \tilde{H}_{\mu\nu}(c_\nu/c_\mu)}{[E - H_{ll}]} \tag{4}$$

for BW. The dependence of denominators on the unknown exact energy eigenvalue E is one of the distinguishing features of Brillouin−Wigner methods. This property is responsible for the existence of a natural gap that may lead to a rapid convergence of perturbation series. Here, $H_{l\mu} = \langle\chi_l^\mu|H|\phi_\mu\rangle$, and $H_{ll} = \langle\chi_l^\mu|H_0|\chi_l^\mu\rangle$. χ_l^μ stands for a general (mono/bi)-excited function from the ϕ_μ. H_0 is the zeroth-order Hamiltonian. In practice, however, for a given MR reference wave function, the MRPT energy depends on the definition of the zero order Hamiltonian. Different variants of MRPT deviate from each other in the selection of these quantities.

As in the SS-MRPT method, the partitioning of H and the treatment of the size-extensivity correction term are independent, and we can even choose $\mathbf{H}_0$ to be a one-particle operator, as that for a truly MP theory. In SS-MRPT, the zeroth-order Hamiltonian, $\mathbf{H}_0$, is different for different CSFs, ϕ_μ or ψ_0 [henceforth, we refer to $\mathbf{H}_0$ for ϕ_μ as $H_{0\mu}$]. In our MP partition, we choose $H_{0\mu}$ to be a sum of the Fock operators for the function ϕ_μ. This will correspond to a multipartitioning MP perturbation theory analogous to what was originally advocated by Malrieu and co-workers.[10] Our SS-MRPT thus utilizes a multipartitioning strategy. We can also envision using an EN type of partition for H. In both the choices, $H_{0\mu}$ is a diagonal operator, and this lends a simpler structure to the SS-MRPT. The SS-MRPT uses the best traits of the multipartitioning strategy as well as of a rigorously size-extensive formulation.

In the cluster amplitude finding equation of the SS-MRPT [see eqs 3 and 4], two sets of terms are present: (i) the first term, called the "direct" term [analogous to the SR part for each ϕ_μ, thus preserving the structure of SR perturbation theory], and (ii) the second part, called the "coupling" term, which is mainly responsible for the twin goals of the formulation: alleviate intruders (which are ubiquitous in a traditional effective Hamiltonian based MR approach) and maintenance of rigorous size extensivity. This plays an important role in the wave function being size-consistent with localized molecular orbitals. It is worth mentioning that the reliability of a computational approach as applied to large systems and chemical reactions depends critically on its ability to maintain the size consistency of calculated properties. Hence, the coupling term(s) have physically and numerically distinct contributions to the electron correlation problem for MR situations. This distinction is the primary feature of the SS-MRPT. The working equations of the SU- and SS-based MRPT methods differ in the form of the coupling terms. However, a repeated solution of essentially the same set of equations for each state is required in the case of SS-based methods, whereas in the conventional SU approach, all states are considered simultaneously.

We note that eqs 3 and 4 involve the coefficients c_μ^0/c_μ explicitly indicating that the cluster amplitudes depend on them. In contrast, coefficients do not appear(s) in the cluster amplitude equation of pure state universal theory. The coefficients and the required energy of the target state are generated by diagonalizing an effective operator (non-Hermitian) $\tilde{H}_{\mu\nu}^{(2)}$ defined in CAS (or CMS):

$$\sum_\nu \tilde{H}_{\mu\nu}^{(2)} c_\nu = E^{(2)} c_\mu \tag{5}$$

Once the cluster amplitudes, $t_\mu^{l(1)}$, are determined using the reference coefficient, the reference function can be updated by diagonalization of the matrix $\tilde{H}_{\mu\nu}^{(2)}$, and thus a self-consistent procedure is obtained. The number of unknowns in this formalism is exactly the same as in the corresponding SUMR-based theory. As the theory is state specific, only one eigenvalue corresponding to the target state represents the exact energy, while the remaining eigenvalue(s) have no physical meaning. That is, if the number of linearly independent functions in the CAS (i.e., Hilbert space) is N, only one root is meaningful (represents the exact intruder free energy); the rest of the roots are extraneous, which is the principal difference with the state-universal-based method(s). It should be emphasized that the strict separability of

the matrix elements of the effective Hamiltonian ensures the separability of its eigenvalues and eigenvectors provided that complete model spaces and localized orbitals are used. Very recently, Evangelista et al.[41] observed that localizing the orbitals to maintain rigorous size consistency of the method does not affect the accuracy of the results obtained.

Equations 3, 4, and 5 are the working expressions used in this work for the evaluation of cluster operators and energy in the perturbative framework. The sets $\{T^\mu\}$ and $\{c_\mu\}$ are coupled through eqs 4 and 5 in the case of BW. From the working equations, it is evident that the c_μ's are updated iteratively for the BW case, whereas c_μ^0 is used for RS scheme to evaluate the cluster amplitudes and $\tilde{H}$. In the RS case, model space combination coefficients get relaxed during the computation of energy when this is obtained by diagonalization. The CAS energy, E_0 and coefficients (unrelaxed) for the reference functions, c_μ^0 's are obtained by diagonalizing the matrix of $H_{\mu\nu}$. Thus, for the RS theory, E_0 (CAS energy) appears in the denominator, while in BW, the target state energy appears. It should be noted that all the variables [c_μ, T^μ] are determined self-consistently for the BW case. Thus, the computation of the SS-MRPT energy requires knowledge of first-order cluster amplitudes which can be obtained via an iterative procedure because of the presence of coupling terms. There are several provisions about how to perform iterations of the cluster equations in the SS-MRPT method.[14] Consideration of different solution techniques of the first order equations is an important aspect. Very recently, Evangelista et al.[41] demonstrated that the iterative solution of the SS-MRPT amplitude equations is not a bottleneck issue. The perturbative energy for the RS scheme is a (up to) second-order quantity, whereas for the BW expansion, this is pseudo-second-order in nature [the energy is no longer rigorously second order] as the amplitude finding equations for BW are not truly first-order due to the presence of explicit $\tilde{H}_{\mu\nu}$ in an amplitude equation, eq 4. In eq 5, there is a choice: one can use it to compute the energy either as an expectation value with respect to the unrelaxed (or frozen) function or by diagonalization in the relaxed form. It should be noted that, in SS-MRPT(RS), the effect of relaxation of the coefficient is somewhat different compared to the traditional MR-MBPT. The effect of large mixing of reference functions and consequent relaxation cannot be estimated fully by SS-MRPT(RS), because frozen c_μ's figure in the amplitude finding equation. SS-MRMPPT(RS) can be regarded as an approximation to the SS-MRMPPT(BW) method. But, we expect that the SS-MRPT(BW) scheme takes this relaxation fully.

Working equations clearly offer the solution to the *intruder state problems* for low-lying electronic states as long as E_0/E is well separated from the external space. This is a convenient aspect of applying the SS-MRPT method to PES, illustrated in the next section. It is noteworthy that, if the contribution of an intruder state is significant, then the "shifting technique to avoid the intruder" will not correct the situation rigorously, and a substantial error in the perturbation energy may be expected. In such situations, one may expand the reference space to include the intruder state. It is well-known that in most cases the effect of intruder states becomes more pronounced away from equilibrium region, and successful treatment of this aspect improves the accuracy of predicted energies.

The development of IMS-based SSMR theory[18] may have the potential to avoid or at least to attenuate the instability of the theory when the virtual determinants do not remain reasonably well-separated in energy from the state energy. To attenuate the computational cost of the CAS-based SS-MRPT method, one can use the IMS-based SS-MRPT scheme. Another way to reduce the computational demand would be to work with a contracted description of the ansatz of the starting wave function as that in the case of the contracted MRCI method.[18]

It should be underlined that mere elimination of small denominators is not sufficient to ensure the convergence of cluster finding equations [eqs 3 and 4] because of the explicit presence of c_μ in it. In the development of the SS-MRPT amplitude finding equations, we have assumed that all the elements of the eigenvector c_μ are nonzero, which allows us to divide the μth cluster equation by c_μ. The working equation is prone to numerical instability owing to the presence of the coefficients c_μ (or c_μ^0) in the denominator, and the paucity of these makes the terms explode. This is an unwarranted situation. This issue is completely different from the conventional intruder effect (related to the vanishing energy difference in the denominator of the cluster finding equations, leading to convergence difficulties). If the reference coefficients go to zero (nearly vanishing), the value of $\langle\chi_l^\mu|T^{\nu(1)}|\phi_\mu\rangle\tilde{H}_{\mu\nu}c_\nu/[\langle\chi_l^\mu|T^{\nu(1)}|\phi_\mu\rangle H_{\mu\nu}c_\nu^0]$ becomes small, akin to the values of the coefficients c_μ. As in any reference function (ψ_0), if the contribution of some of its component function (ϕ_μ) is very small, then the back coupling from other components in the cluster finding equation should not be large, at least not larger than the zero-order value; i.e., $\langle\chi_l^\mu|T^\nu|\phi_\mu\rangle H_{\mu\nu}c_\nu$ will be as small as c_μ. Consequently, the problem of divergence is alleviated. Practical calculations show that SS-MRPT method has good convergence properties even for molecules far away from their equilibrium geometries. In our present applications, we have observed that the amplitude finding equations do not suffer from numerical instability even when all coefficients (c_μ) have been included. But still, the appearance of c_μ in the denominator invites a threat for the calculations of excited states. Recently, a numerically more robust approach has been suggested and implemented by Evangelista et al.[41,48] Explicit discussion of this issue has appeared in the recent paper of Engels-Putzka and Hanrath.[49] In this paper, they devised a technique which is very effective in drastically reducing the numerical instability of amplitude finding equations without an undue sacrifice in accuracy. In passing, we should mention that the cluster finding equations in the SUMR[37] and BWMR[12,50] approaches have no c_μ coefficient in the denominator and are free from such difficulties.

In this article, we have considered MP type partitionings. In our numerical application, we used the following definition of the Fock operator in the case of SS-MRMPPT:

$$f_\mu = \sum_{ij}\left[f_{\text{core}}^{ij} + \sum_u\left(V_{iu}^{ju} - \frac{1}{2}V_{iu}^{uj}\right)D_{uu}^{\mu}\right]\{E_i^j\} \qquad (6)$$

Here, u represents both a doubly occupied and a singly occupied active orbital in ϕ_μ, and D^μ's are the densities labeled by the active orbitals. Since our H_0 is always diagonal for MP scheme, the zeroth-order Hamiltonian operator is $H_0^\mu = \Sigma_i f_\mu^{ii}\{E_i^i\}$. Thus, H_0 is built on one-particle operators; i.e., it is monoelectronic in nature. The use of diagonal zero-order operators ensures the maximum computational simplicity of the scheme, which can attenuate computational cost in rather complicated situations. For the sake of comparison, results of the EN partition (appropriately generalized in the context of multipartitioning[10]) have also been presented.

This discussion will remain incomplete if we do not discuss the characteristic features of the SS-MRPT with respect to the multireference state specific Brillouin−Wigner second-order perturbation theory (BWPT) of Hubać and co-workers.[12] In this part, we will analyze qualitatively the structural kinship and scaling (size-extensivity) property of these two methods. Theoretical derivations and detailed discussions of further aspects of these approaches are given elsewhere.[12,14] A complete active space is employed as a reference for both the perturbation theory studies. Here, we want to state that both the BWPT of Hubać and co-workers[12] and the SS-MRPT developed by Mukherjee and co-workers[14] are based on the "diagonalize then perturb" philosophy.

In the state-specific Brillouin−Wigner PT of Hubać and co-workers,[12] working equations are obtained by introducing the SS wave operator in the BW form of the Bloch equation. The equations that determine the amplitudes of the wave operator are coupled only through the exact energy of the target state (the price paid is the lack of size extensivity of the method), a computational advantage emphasized by Hubać and co-workers. Thus, in BWPT, the amplitude equations naturally decouple for each reference and do not contain expansion coefficients for the reference wave function. On the other hand, the development of the SS-MRPT formalism of Mukherjee and co-workers involves inserting the state-specific JM wave operator into the expansion for the exact wave function and using the physically motivated sufficiency condition discover cluster finding equations. As a result of this, in contrast to the BWPT, the amplitudes of the SS-MRPT for reference μ are dependent on the amplitudes of the entire model space through the renormalization (or coupling) terms. We recall that, in the SS-MRPT formalism, the renormalization terms have been exploited to attain the twin goals of ensuring size extensivity and avoiding intruders. Although both the theories of Mukherjee and co-workers and of Hubać and his group use sufficiency conditions, it is worth noticing the difference in the sufficiency conditions of BWPT and SS-MRPT. In contrast to the BWPT approach, the SS-MRPT(BW) method explicitly contains the eigenvector coefficients. This reference coefficient weighing originates from the sufficiency condition used.

To make some comment on the formulation of BWPT proposed by Hubać and co-workers,[12] we have written the perturbative equations with the same $\mathbf{H}_0$ used by us and with the energy parameter unexpanded. Their amplitude equations, in place of the coupling term, contain the target energy E itself, multiplied by certain expressions containing just the amplitudes of the $\mathbf{T}^\mu$. The coupling of the various t_i^μ's with different μ's is thus implicit, appearing via E, since E involves all the cluster operators with different μ's. The equation for the first order cluster operator in our terminology is

$$\langle\chi_l^\mu|H|\phi_\mu\rangle + \langle\chi_l^\mu|H_0T^{\mu(1)}|\phi_\mu\rangle = E\langle\chi_l^\mu|T^{\mu(1)}|\phi_\mu\rangle \qquad (7)$$

Using the multipartitioning expression[10] of H we get

$$\langle\chi_l^\mu|H|\phi_\mu\rangle + (H_{ll} - E)\langle\chi_l^\mu|T^{\mu(1)}|\phi_\mu\rangle + \sum_{m\neq 1}{}' \langle\chi_l^\mu|[H_0]_\mu|\chi_m^\mu\rangle\langle\chi_m^\mu|T^{\mu(1)}|\phi_\mu\rangle = 0 \qquad (8)$$

A strong objection to the choice of sufficiency conditions used by Hubać and co-workers to yield amplitude equations is that BWPT is not rigorously size-extensive due to the presence of disconnected terms stemming from $E\langle\chi_l|T^{\mu(1)}|\phi_\mu\rangle$. There are no counter terms to cancel this disconnected contribution. The remaining terms of this equation are connected in nature. Attempts to restore size extensivity in the formalism of BWPT attempt to do so by way of expanding the target energy E in terms of an unperturbed RS-like energy E_0. The size-extensivity correction in this way does not necessarily improve the quality of results with respect to the nonextensive parent theory.[51] It has been observed that the intruders would not show up if the inextensivity correction is incorporated by one iteration only. However, in general, this procedure does not ensure the removal of all the inextensive terms. Multiple iterations or the converged RS type of results will, however, unfortunately bring back the problem of potential intruders. In spite of that, recent calculations have shown that this approach produces very promising results in some difficult quasi-degenerate situations. On the other hand, SS-MRPT(BW) is both size-extensive and size-consistent (when localized orbitals are used). The cluster amplitudes in the BW series are generally obtained in the SS-MRPT theory by using BW-type denominators but are corrected for size inextensivity with counter terms originating from coupling terms; such counter terms are inherently absent in the BWPT approach.[12] In SS-MRPT(BW), there is an extra term $\Sigma_\nu\langle\chi_l|T^{\mu(1)}|\phi_\mu\rangle H_{\mu\nu}(c_\nu/c_\mu)$ which cancels the disconnected term; as a result, the final cluster finding equation is manifestly connected. Hence, the amplitude finding equations in SS-MRPT involve an explicit coupling between the cluster operators for all the μ's as demanded by the rigorous requirement of size extensivity. Proving the extensivity and consistency of both cluster amplitudes and the energy E of SS-MRPT is a rather involved exercise, and we refer to the original papers for details.[14,15]

The above-mentioned, nice formal properties provide a unique niche for the SS-MRPT approach that has opened the possibility toward accurate treatment for the electronic states, especially when investigating large systems that are computationally intractable for the more robust electron correlation methods. The SS-MRPT method resolves the contradiction between the conditions for a reliable as well as good convergence and those for asymptotic separability in the context of the MRPT approach.[10] The goodness of

the SS-MRPT method in different partition and expansion schemes will be examined by way of example calculations in the next section, where we shall illustrate the accuracy of the SS-MRPT method for a number of homonuclear diatomic halogen and BH molecules. The performances of the SS-MRMPPT(RS) and SS-MRMPPT(BW) methods based on the same CAS have been compared. To validate our numerical implementation, we tested our results against independent numerical results available in the literature. The remainder of this paper consists of discussion of our numerical applications.

III. Numerical Applications: Dissociation PES of Molecular Systems

In the present paper, we deal with the dissociation of single bond systems. Despite their geometric simplicity, the electronic structure of these molecules provides a difficult challenge for MRPT. We have organized the sequence of the presentation of our numerical results in two parts: (A) dissociation of X_2 [X = F, Cl, and Br] systems and (B) the calculation of the ground state PES of the astrophysically important BH molecule in order to show the applicability of the SS-MRPT approach to a open-shell case. In this paper, we have also investigated the effects of orbital rotation on the total SS-MRMPPT energy.

In the approach pursued here, the recovery of correlation is perceived as a two-stage process. The procedure we used involved running an initial CASSCF calculation followed by SS-MRMP analysis. The SS-MRPT correction ensures (provides the second order energy in the full active space) a proper treatment of the dynamic electron correlation in the wave function. All current calculations use CASSCF for the description of the reference state because it is size-consistent and correctly describes the static correlation. CASSCF wave functions are especially useful for exploring the details of PES. In the CASSCF method, the active electrons are distributed in all possible ways over the active orbitals with a given space-spin symmetry of the state considered. The active space provided by the user of the CASSCF software represents a key point to obtain accurate theoretical predictions, once the dynamic correlation has subsequently been taken into account, for instance, at the SS-MRPT level. In our CASSCF calculations for electronic states of the X_2 and BH systems, two electrons were active, and the active space included two orbitals [denoted as CAS (2,2)] for each internuclear distance. In our calculations of the BH system, we also used the same CAS, termed CAS(val), as the one reported in ref 52, and thereby the performance of the various SS-MRMPPT variants can be assessed by comparing our results with the results reported in ref 52 of Sherrill and coworkers. The choice of the active space stemmed from the electron configuration of the ground state of the molecules. In order of increasing severity, our tests include the homonuclear diatomic molecules F_2, Cl_2, and Br_2. The single bonding in X_2 is basically a chemical reaction that involves one bonding orbital(s) and the corresponding higher-lying antibonding orbital. CAS(2,2) is thus the smallest active space that allows for a qualitatively correct treatment of the bond breaking since, in the dissociation region, both bonding and antibonding orbitals become quasi-degenerate nonbonding orbitals. On the other hand, the larger reference space provides improved first order energies and smaller overall perturbation corrections. Not only that, increasing size of the reference space enhances the diagonal perturbation matrix elements. It might be expected that the perturbation series displays slow convergence with the larger reference spaces in comparison to the smaller one. It is always desirable to use as accurate a description of the unperturbed state as possible. In this computational method, the full reaction coordinate is computed at the CASSCF level of theory [using GAMESS(US) quantum-chemistry software]. For the perturbation calculations, mono- and bielectronic integrals are calculated by the GAMESS(US) program package.

A. Ground State PES of Homonuclear Dihalogen Molecules X_2. Theoretical studies on ground electronic states along the dissociation path of X_2 are very challenging as it possesses quasi-degeneracy at some point on the reaction path and there are potential intruders at some other points in the PES. This is the main reason (to our knowledge) for the inapplicability of state-universal MRCC theory to generate spectroscopic constants of X_2. Thus, these systems are appropriate to test the efficacy of any state-specific MR-based theory. Therefore, in order to obtain a PES accurate enough in the whole range of the reaction path, computation of both nondynamical and dynamical electron correlation in a sophisticated and intruder free manner is very much essential.

Before investigating the properties of Cl_2 and Br_2, we consider the demanding example of F_2, which is well-known to have multireference character in its equilibrium description. The fluorine molecule is one of the most widely studied and still one of the most difficult diatomic molecules in terms of obtaining a correct dissociation behavior. Almost any new single or multiconfiguration reference approach to the correlation energy has been long-since checked for the computation of the lowest state of the F_2 molecule as a simple multireference system that is difficult to solve.[14,26,45,53−60] A proper description of the reference wave function of equilibrium and the entire reaction coordinate including dissociated X_2 requires a linear combination of two closed-shell determinants: (core)σ_g^2 and (core)σ_u^2 in the D_{2h} point group. As the stretching of the X_2 bond increases, the contribution from the second determinant to the total CASSCF(2,2) wave function increases significantly. Needless to say, the SR-based method noticeably underestimates the second determinant contribution. Recent papers by Pittner et al.[26] and Evangelista et al.[48] provide a nice summary of the performance of various many-body methods applied to the F_2 problem. A comprehensive tabulation of spectroscopic constants of F_2 obtained from different perturbation theory have been found in a paper by Rosta and Surjan.[59] The dissociation energy of F_2 has been determined by Yang et al.[61] using ion-pair dissociation imaging. In very recent articles, Bartlett and Musial[62] have given a detailed discussion of the performance of a new hierarchy of SR-based coupled-cluster methods, *nCC* for the F_2 bond-breaking process. Very recently, Evangelista et al.[48] published a nice paper regarding the performance of various MRCC methods to calculate the dissociation energy of F_2 in considerable

detail. It should be emphasized here that the most extensive application at the production level of SS-MRCC theory by Mukherjee and co-workers[39] beyond double excitation was carried out by Evangelista et al.[48] [termed by them as Mk-MRCC]. They provide a benchmark for other high-accuracy calculations of the F_2 dissociation surface. In this paper,[48] they stated that the low accuracy of MR-BWCCSD theory to describe the dissociation reaction of F_2 in comparison to the Mk-MRCCSD is due to the size inextensivity of BWCC method. Recently, Ruedenberg and co-workers[63] established the nonrelativistic PES taking into account electron correlations only in the valence shell, while to get very accurate results, one must also include the effects of relativity. The spectroscopic results of Evangelista et al.[48] along with the present work allow an assessment of effectiveness of the SS-MRMPPT in comparison to the computationally demanding parent, full-blown SS-MRCC (Mk-MRCC) method. It is evident from the above discussion that F_2 provides a unique testing ground for different theoretical approaches to the study of PESs. Because the F_2 system is small enough for the application of very large basis sets close to the limit, we have done a series of calculations for the system to get a definite answer to the performance of the SS-MRMPPT with the RS and BW scheme. The results for F_2 have been obtained with a variety of basis sets. The perturbation series converges safely if diffuse basis functions are added to the basis set. It is worth mentioning that, in our numerical application, we have not observed any unphysical kink (or barrier) in the F_2 dissociation PES, but Mášik et al.[57] have observed that there is an unphysical barrier in the CASPT2 dissociation surfaces near $R(FF) = 3$ Å.

For our computations on the F_2 system, let us first consider the frozen-core calculations using the DZP+basis[54] [which is the standard Huzinaga−Dunning DZ set with the most diffuse p function uncontracted and augmented by six Cartesian d functions ($\alpha_d(F) = 1.580$)] where we have various theoretical results[26,45,53−60] for comparison. Spectroscopic constants of F_2 at various levels of theory are available in the literature for this basis. Hence, we have performed our calculations using this basis to demonstrate the efficacy of the SS-MRPT for different partitions and expansion schemes. The spectroscopic properties (equilibrium bond length and dissociation energy) derived from our PESs are reported in Table 1 for the F_2 molecule using the DZP+ basis. The fluorine 1s core orbitals have been kept frozen (uncorrelated) in our calculations. Comparisons are presented with other multireference perturbation calculations focusing only on the relative performance of the methods of Hirao (MRMPPT),[3] Roos and co-workers (CASPT2),[5] and the APSG (antisymmetrized product of strongly orthogonal geminals) PT of Surján and co-workers.[59] For the sake of comparison, we have also summarized the results of spin-flip and MR-CISD methods.[58] For a balanced comparison, we also report in Table 1 the best nonrelativistic, valence-correlated, CBS results from Bytautas et al.[63] The FCI result is not available at this level, and the results from Bytautas et al. have been taken as a reference. The results in the table show the ability of both versions of SS-MRMPPT to reproduce the established theoretical values well. The results from the CASSCF, GVB (generalized valence bond), and APSG (represents an independent electron-pair model) calculations compare well with spin-flip-SCF (SF-SCF). Although incorporation of the dynamical correlation in the SF-SCF method improves the results, the SS-MRMPPT represents a somewhat better description for the spectroscopic constants than SF-SCF based correlated methods [such as SF-CIS(D), SF-OD, and so on]. From the table, it is abundantly clear that the agreement of the SS-MRMPPT results with state-of-the-art *ab initio* results of Bytautas et al.[63] (and also the experimental ones) is definitely better than for CASPT2, GVB+HPT, and APSG+HPT and is of the same quality as the MRMPPT of Hirao.[3] It seems therefore reasonable to say in this case that SS-MRMPPT is not only a useful MRPT approach but also a strong competitor to other well established MRPT methods such as CASPT2, MRMPPT, and so on. A major advantage of the computed SS-MRMPPT energy is that it is rigorously size-extensive in nature in contrast to the MRMPPT and CASPT2. It may be of interest to note that Rintelman et al. investigated extensively the size extensivity problem of MRMPPT and CASPT2 with the F_2 molecule using a series of basis sets. To illustrate that this high accuracy of the SS-MRMPPT in predicting spectroscopic constants is not an accident, we re-evaluate the same using various correlation-consistent basis sets.

The DZP+basis is too small to illustrate a sensible comparison of the computed spectroscopic data with the corresponding experimental results. Actually, in the case of F_2, the use of a basis set up to quadruple-ζ quality is sensible (in accordance with the literature) in order to demonstrate the effect of the basis set and to compare the calculated results with the experimental data. In Figures 1 and 2, we

Table 1. Spectroscopic Constants for the Electronic Ground State of F_2 Molecule Using DZP+ Basis Sets

method	R_e (Å)	D_e (eV)
CASSCF	1.5107	0.63
SS-MRMPPT(RS)	1.4330	1.42
SS-MRMPPT(BW)	1.4321	1.35
GVB[a]	1.512	0.453
APSG[a]	1.501	0.486
MRMPPT2[a]	1.4254	1.440
GVB1H1PT[a]	1.445	1.025
GVB1H2PT[a]	1.420	1.180
APSG1H1PT[a]	1.447	1.026
APSG1H2PT[a]	1.423	1.274
MRCISD10[b]	1.435	1.222
MRLCCM10[b]	1.439	1.221
MRCISD32[b]	1.436	1.275
MRLCCM32[b]	1.439	1.257
CASPT2[c]	1.442	1.275
MRMBPT(2)[c]		
simple averaging	1.374	2.348
double averaging	1.377	2.196
SF-SCF[d]	1.567	0.28
SF-CIS(D)[d]	1.429	1.14
SF-OD[d]	1.437	1.24
VOO-CCD(2)[d]	1.417	1.51
MR-CISD[d]	1.435	1.22
best *ab initio*[e]	1.4148	1.70
experiment	1.412	1.66

[a] Ref 59. [b] Ref 54. [c] Ref 57. [d] Ref 58. [e] Ref 63 (2,2) CAS has been used in our works.

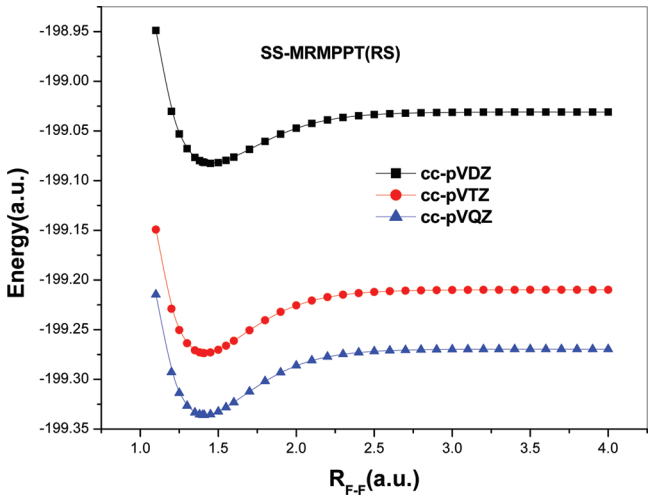


Figure 1. Potential energy surfaces for the F_2 molecule with the SS-MRMPPT(RS) method in different cc-pVXZ basis sets.

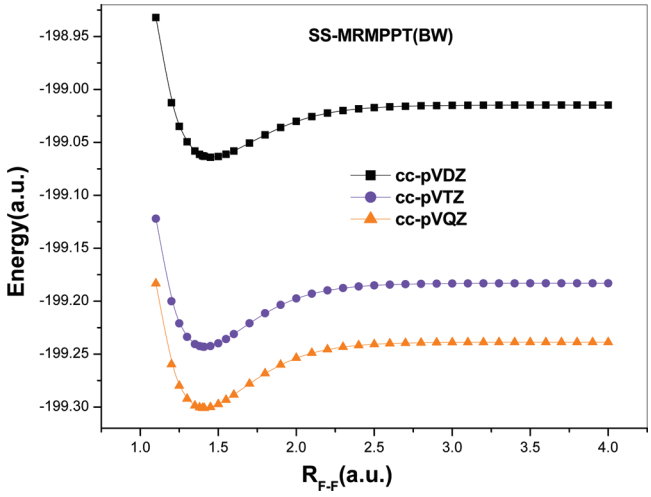


Figure 2. Potential energy surfaces for the F_2 molecule with the SS-MRMPPT(BW) method in different cc-pVXZ basis sets.

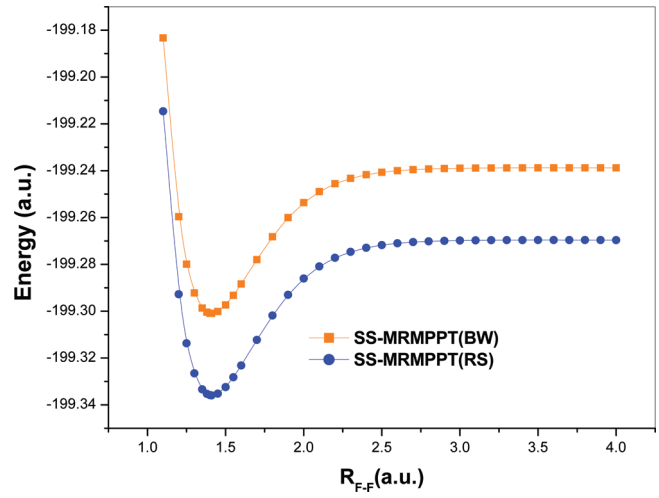


Figure 3. Potential energy surfaces for the F_2 molecule with the SS-MRMPPT(RS) and SS-MRMPPT(BW) methods in the cc-pVQZ basis set.

draw the SS-MRMPPT PESs for F_2 using a different correlation consistent basis sets. From the PESs, it can be seen that a qualitatively balanced description is obtained by SS-MRMPPT approaches for various basis sets. From our computations, we have observed that the SS-MRMPPT(BW) potential surfaces are situated above the SS-MRMPPT(RS) surfaces at all geometries for all basis sets considered here [see Figure 3]. As the traditional full configuration interaction (FCI) and the high level PES calculations are not available for these basis sets, comparison of SS-MRMPPT methods is made with the spectroscopic parameters of high level methods reported in the literature, and thereby one can envision the correct description of bond breaking. Table 2 contains the spectroscopic constants for F_2 in various cc-pVXZ basis sets with different SS-MRPT methods using EN and MP partition schemes. These basis sets would be expected to describe, for example, the anion character of the F^+F^- charge transfer components of the wave function with more flexibility. It is interesting to compare our results with the Mk-MRMP2 results of Evangelista et al.[41] In Table 2, we present results of the CASSCF and CCSD as well as CCSDT for comparison. Further, we tabulate a variety of state selective equation of motion coupled cluster based methods of Nooijen.[55] Results of auxiliary-field quantum Monte Carlo method (AFQMC) calculations, which yields spin-contamination-free results within a Hilbert space, of Purwanto et al.[60] have also been reported in Table 2. BW-MRCCSD results and the rather recent results from the Mk-MRCC method of Evangelista et al.[48] are also given in the same table to calibrate the quality of our results because the Mk-MRCC method serves as a benchmark. In fact, methods such as Mk-MRCC and BW-MRCC provide useful results for our understanding of the electronic structure of molecules despite their high computational cost. Due to this high numerical cost, applications of such methods remain constrained to relatively small systems in contrast to the corresponding MRPT part. We observed that SS-MRMPPT and Mk-MRPT2 methods are of comparable accuracy. Here, we recall once more that, in the Mk-MRPT2 computation, Evangelista et al. used unrelaxed description, and we have already discussed the importance of the issue of relaxed variants. Table 2 displays that the performance of the SS-MRMPPT is better than the CCSD and various EOM-CC methods. It is found that, even though only a single bond is broken in the dissociation considered here, nonetheless, sextuple excitations with respect to single-determinant reference functions are required to recover the binding energy with a "chemical accuracy". This view is supported by the fact that dissociation energy and vibrational frequency provided by CCSD are not as good as the other methods reported in the table. Recently, Purwanto et al.[60] observed that, in the case of F_2 with the cc-pVXZ basis, the RCCSD(T) method breaks down in the dissociation limit. On the other hand, the PES provided by UCCSD(T) is correct in the dissociation limit, but the shape of the surface near the equilibrium distance is distorted in nature, which invites significant error in the intermediate geometries. From the table, we have observed that the SS-MRMPPT results for each basis set are comparable to those obtained with the AFQMC method. Here, it is important to note that the

Table 2. Spectroscopic Constants for the Electronic Ground State of F_2 Molecule Using Different cc-pVXZ Basis

basis	method	R_e (Å)	ω_e (cm^{-1})	D_e (eV)
cc-pVDZ	CASSCF	1.5309	575	
	SS-MRMPPT(RS)	1.4546	809	1.40
	SS-MRMPPT(BW)	1.4557	795	1.34
	SS-MRENPT(RS)	1.4342	887	2.15
	SS-MRENPT(BW)	1.4365	865	1.89
	Mk-MRPT2[a]	1.4454	815	1.29
	CCSD[b]	1.432	885	0.96
	CCSDT[b]	1.4577	787	1.18
	BW-MRCCSD[b]	1.4469	821	1.63
	Mk-MRCCSD[b]	1.4548	793	1.37
	AFQMC[c]	1.467	725	1.29
	RCCSD(T)[c]	1.4751	785	1.18
	UCCSD(T)[c]	1.4428	853	1.14
	LSDA[c]	1.3970	1026	3.45
	B3LYP[c]	1.4097	1033	1.64
cc-pVTZ	CASSCF	1.4696	699	0.72
	SS-MRMPPT(RS)	1.4151	918	1.71
	SS-MRMPPT(BW)	1.4134	917	1.62
	SS-MRENPT(RS)	1.4001	1017	2.63
	SS-MRENPT(BW)	1.3984	987	2.26
	Mk-MRPT2[a]	1.4055	955	1.61
	CCSD[b]	1.3946	1012	1.22
	CCSDT[b]	1.4154	923	1.50
	BW-MRCCSD[b]	1.4060	953	2.03
	Mk-MRCCSD[b]	1.4127	925	1.71
	AFQMC[c]	1.411	928	1.70
	RCCSD(T)[c]	1.4131	926	
	UCCSD(T)[c]	1.3987	1022	1.49
	LSDA[c]	1.3863	1065	3.49
	B3LYP[c]	1.3957	1072	1.65
	DIP-STEOM-CCSD[2-][d]	1.4353	834	
	DIP-EOM-CCSD[2-][d]	1.4249	811	
	CSS-EOM-CCSD[B][d]	1.4173	901	
	RSS-EOM-CCSD[B][d]	1.4126	927	
cc-pVQZ	CASSCF	1.4688	696	0.72
	SS-MRMPPT(RS)	1.4120	931	1.80
	SS-MRMPPT(BW)	1.4094	921	1.69
	SS-MRENPT(RS)	1.3994	1016	2.71
	SS-MRENPT(BW)	1.3962	988	2.31
	Mk-MRPT2[a]	1.4023	958	1.68
	CCSD[b]	1.3906	1016	1.29
	CCSDT[b]	1.4124	925	1.58
	BW-MRCCSD[b]	1.4024	955	2.12
	Mk-MRCCSD[b]	1.4093	926	1.79
	AFQMC[c]	1.411	912	1.77
	RCCSD(T)[c]	1.4108	929	
	UCCSD(T)[c]	1.3946	1036	1.567
	LSDA[c]	1.3856	1062	3.47
	B3LYP[c]	1.3944	1109	1.63
	best *ab initio*[e]	1.4148	920	1.70
	experiment	1.412	916.64	1.66

[a] Ref 41. [b] Ref 48 [for symmetry-adapted natural orbitals]. [c] Ref 60. [d] Ref 55. [e] Ref 63. Experiment: ref 86 (2,2) CAS has been used in our works.

CAS(10,12) has been used in the AFQMC method, whereas in our calculation we have used CAS(2,2). Computationally, the SS-MRPT method is less demanding than the AFQMC one. The values of spectroscopic constants of the SS-MRMPPT in Table 2 are in better agreement with the highly accurate available *ab initio* calculation[63] (and also experiment) than those of the LSDA and B3LYP methods. It should be noted that the shapes of the LSDA and B3LYP PESs are not correct in the intermediate region of UCCSD(T) in contrast to the SS-MRMPPT one. As seen in Table 2, the errors in the computed spectroscopic constants are usually large for several of the methods for the cc-pVDZ basis relative to that of the other basis. It is important to realize, however, that the poor performance of various methods for cc-pVDZ basis is not in its inability to describe charge transfer components of the wave function, but rather it is an inflexibility of the valence region itself. It has been found that, in order to obtain good results, the use of quite large and proper atomic orbital basis sets is clearly desirable. Analyzing the results in Table 2, we find that the calculated spectroscopic constants of the ground state via different variants of SS-MRPT agree reasonably well with the corresponding experimental values. Our spectroscopic constants agree quite well with the previously calculated theoretical values. In general, the SS-MRMPPT provides superior results (for the spectroscopic constants of F_2) relative to the BW-MRCC methods reported by Evangelista et al.[48] The present calculations of F_2 produce almost the same equilibrium distance, vibrational frequency, and dissociation energy as those reported in Mk-MRCC. In other words, the SS-MRPT results are competitive with the Mk-MRCC to deliver accurate energetics for the dissociation of F_2 but at lesser cost in the calculation. It is worth noting that the SS-MRPT method is computationally cost effective for realistic applications, and it is able to achieve a similar level of accuracy as the Mk-MRCC method for the treatment of the dynamical correlation and can thus serve as a reference for calibration of more approximate approaches. Table 2 clearly displays that the various SS-MRMPPT results do not differ significantly in the general trends for a given partitioning. In this work, we have also listed in Table 3 the comparative performance of the various methods reported by Shepard et al.[64] for the sake of comparison. Our observations are that the SS-MRMPPT results generate spectroscopic properties that are better than the best single-reference MP methods summarized in this study (MP3, MP4). From Tables 2 and 3, we have observed, in general, our current multireference calculations via SS-MRMPPT methods predict R_e values with an accuracy competitive to, or better than, the MR-SDCI and MR-AQCC methods, despite the fact that the MR-AQCC method is computationally costly. In contrast to the SS-MRMPPT method, the calculated R_e values using the MR-CISD and MR-AQCC methods tend to be smooth and monotonic with basis set improvement. From Tables 2 and 3, as far as the prediction of spectroscopic constants is concerned, we note that the general performance of the SS-MRMPPT is better than its EN counterpart. Being a symmetrical system, it is well known that the EN partition meets with particular difficulties in the description of the dissociative part of the PES[65] with a delocalized basis of molecular orbitals, as unphysical Coulombic terms and other artifacts appear in the interaction energy of the fragments. In particular, we want to mention that Angeli et al.[66] also observed the defects of the EN partition with respect to dissociation in the case of F_2 even in the case of CAS containing 78 432 determinants [e.g., Figure 1 of ref 66]. Our numerical experience with the EN partitioning is mingled. Very satisfactory results with EN for correlation energies have been obtained in some cases,[18] while in other cases, an overestimation of low-order contributions and very

Table 3. Equilibrium Bond Length for the Electronic Ground State of F_2 Molecule Using Different cc-pVXZ Basis[a]

basis	MP2	MP3	MP4	MRSDCI	MRAQCC	SS-MRMPPT(RS)	SS-MRMPPT(BW)
cc-pVDZ	1.4239	1.4168	1.4504	1.4652	1.4620	1.4546	1.4557
cc-pVTZ	1.3958	1.3837	1.4151	1.4191	1.4184	1.4151	1.4134
cc-pVQZ	1.3275	1.3812	1.4149	1.4153	1.4153	1.4120	1.4094

[a] Results are taken from: ref 64.

Table 4. Spectroscopic Constants for the Electronic Ground State of F_2 Molecule Using cc-pVQZ Basis for Various NEVPT2 along with the SS-MRMPPT Method

method	R_e (Å)	D_e (eV)
SS-MRMPPT(RS)	1.4120	1.80
SS-MRMPPT(BW)	1.4094	1.69
NEVPT2[a]	1.3960	1.717
PCNEVPT2[a]	1.3960	1.720
NEVPT3[a]	1.4171	1.390
FDD-MR(C)/NEVPT2[a]	1.4050	1.751
FDD-MR(C)/NEVPT3[a]	1.4200	1.395
MP-EN[a]	1.4108	1.77
$(SC)^2$2SDCI[a]	1.4129	1.59
MRCI[a]	1.4119	1.62
8-ref CASSCF[b]	1.4076	1.64
8-ref CI[b]	1.4076	1.66
best *ab initio*[c]	1.4148	1.70
experiment	1.412	1.66

[a] Ref 42. [b] Ref 67. [c] Ref 63, (2,2) CAS has been used in our works.

slow convergence have been reported. However, more extensive analysis of several systems is required before one can assume this to be a general conclusion. Work towards this direction is in progress in our laboratory. All of this underlines the importance of a detailed study of the partitioning problem in SS-MRPT, just as with that of the other MRPT. This paper illustrates the considerations that are necessary when choosing a zeroth-order Hamiltonian for the MRPT method. Inspecting Table 2, one observes that in the case of the F_2 molecule with the cc-pVTZ basis, the SS-MRMPPT method yields accurate predictions of the equilibrium distance, vibrational frequency, and dissociation energy.

For the sake of completeness of our comparative study, in Table 4, we have also presented the SS-MRMPPT results along with another well studied internally contracted state-specific MRPT method, variants of the NEVPT2 method[42] which also confirm what is said for the potentiality of the SS-MRMPPT scheme. In a paper by Angeli et al.,[42] various active spaces have been considered. In this table, a selection of the previously published theoretical values of Lourderaj et al.[67] has also been reported. From Table 4, it is observed that the SS-MRMPPT describes the ground state spectroscopic properties to a reasonable accuracy in comparison with the various NEVPT2 schemes, MRCI, $(SD)^2$SDCI, and MP-EN methods, and with respect to the experiments. In some cases, the agreement of the SS-MRMPPT results with highly accurate *ab initio* results[63] is acceptably closer than that for other methods reported in the table. The comparative demonstration in Table 4 clearly illustrates again the efficacy and reliability of the SS-MRMPPT method to predict the spectroscopic properties.

It is interesting to compare our findings with the recently introduced method of correlation energy extrapolation by intrinsic scaling (CEEIS).[63] Table 5 compares the spectroscopic constants relating to the F_2 molecule that are obtained from the CEEIS method of Ruedenberg and co-workers[63] and various sources reported by them in addition to our SS-MRMPPT/cc-pVQZ method. Comparing the results, we find that SS-MRMPPT(RS) and SS-MRMPPT(BW) are both accurate and of comparable accuracy with CEEIS. In this context, it should be kept in mind that the present SS-MRMPPT calculations do not incorporate the correction due to core−electron correlations and relativity effects (including spin−orbit coupling) in contrast to the CEEIS. These incorporations significantly improve the overall quality of the PES and hence can considerably change the values of the spectroscopic constants.[63,68]

Table 5. Spectroscopic Constants for the Electronic Ground State of F_2 Molecule[a]

method	R_e (Å)	ω_e (cm^{-1})	D_e (eV)
SS-MRMPPT(RS)	1.4120	931	1.80
SS-MRMPPT(BW)	1.4094	921	1.69
CEEIS	1.4135	915	1.66
CC-5/R12	1.4122	918.9	
ic-MRCI+Q	1.4105	916.9	1.59
ic-MRCI		899.7	1.49
CASPT3	1.4091	920.0	1.47
(mv)td-2	1.4118	915.2	1.59
4R RMR CCSD(T)	1.416	911.2	
experiment	1.412	916.64	1.66

[a] Various theoretical results are taken from ref 63 [for a detailed discussion, see Tables X and XI of ref 63]. (2,2) CAS has been used in our works.

Here, we also want to mention that (2,2)MR-MP2 calculations done by Barbosa and Barcelos[44] with an aug-cc-pVTZ basis set have given values $R_e = 1.428$ Å, $\omega_e = 888$ cm^{-1}, and $D_e = 1.59$ eV. Malrieu et al.[69] have performed $(SC)^2$SDCI calculations of F2 with a basis set of 5s4p3d2f1g quality and reported $R_e = 1.389$ Å, $\omega_e = 1004.2$ cm^{-1}, and $D_e = 3.05$. In the same paper, Malrieu et al. reported the corresponding values for the (MR-SDCI + Q) method as 1.417, 930, and 1.658, respectively.

Very recently, Zhang et al.[70] have implemented the idea of a locally contracted configuration interaction of singles and doubles (by introducing a coupled-electron pair approximation, CEPA-3) including the leading part of the triples and quadruples in the evaluation of equilibrium bond lengths (A^0) and harmonic frequencies (cm^{-1}) of the F_2 molecule with the TZVP basis set. The results obtained therefrom reveal that the bond length and the frequency respectively vary as (i) LC-CEPA-3, 1.380 and 1052.6; (ii) LC-CEPA-3+TQ/col/line, 1.416 and 911.5; and (iii) LC-CEPA-3+TQ/explicit, 1.415 and 907.8. Our overall observation is that the SS-MRMPPT approach is very competitive in accuracy with the theoretical results of Zhang et al.[70] as well in the case of the F_2 system.

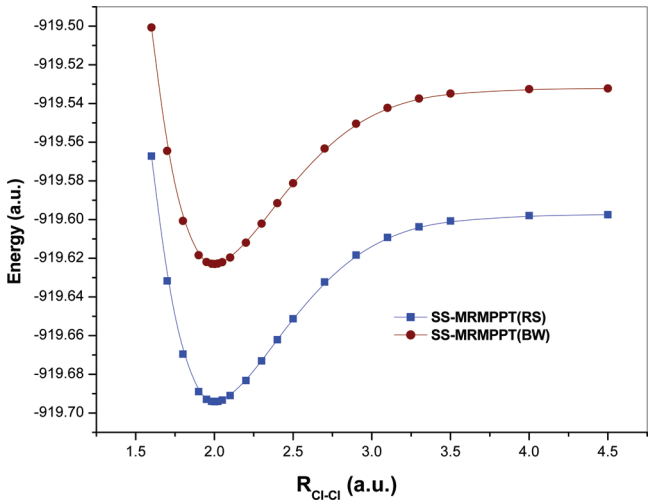


Figure 4. Potential energy surfaces for the Cl_2 molecule with the SS-MRMPPT(RS) and SS-MRMPPT(BW) methods in the cc-pVQZ basis set.

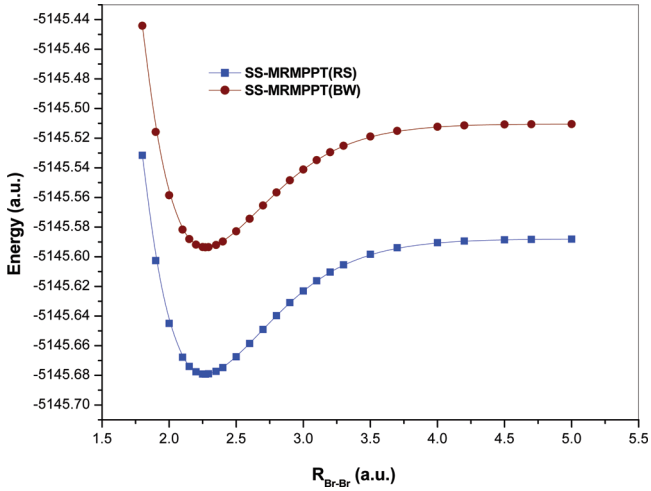


Figure 5. Potential energy surfaces for the Br_2 molecule with the SS-MRMPPT(RS) and SS-MRMPPT(BW) methods in the cc-pVQZ basis set.

Several comparisons of the results of the SS-MRMPPT with respect to the different theoretical calculations including the most accurate available *ab initio* calculations in the context of F_2 molecule have shown that the method is a very useful companion perturbation theory to the parent full blown SS-MRCC(Mk-MRCC) one and can be used to study larger multireference systems, for which SS-MRCC is generally not applicable. It is justified to use the SS-MRMPPT formalism as an effective and acceptable compromise between the computational demands and accuracy, and this we propose to adapt in our next applications (say, Cl_2 and Br_2).

The general properties followed by Cl_2 and Br_2 molecules are very close to each other and can be discussed together. In Figures 4 and 5, we draw the potential surfaces calculated by the SS-MRMPPT methods for Cl_2 and Br_2 using cc-pVQZ basis sets to display the pattern of the computed PESs. The spectroscopic properties using computed PESs at each level of SS-MRMPPT calculation for these molecules along with the experimental results are presented in Tables 6 and 7. The

Table 6. Spectroscopic Constants for the Electronic Ground State of the Cl_2 Molecule

Basis	Method	R_e (Å)	ω_e (cm^{-1})	D_e (eV)
cc-pVDZ	CASSCF	2.0628	468	1.42
	SS-MRMPPT(RS)	2.0367	543	2.02
	SS-MRMPPT(BW)	2.0366	543	1.94
cc-pVTZ	CASSCF	2.0368	504	1.69
	SS-MRMPPT(RS)	2.0085	574	2.51
	SS-MRMPPT(BW)	2.0054	563	2.37
cc-pVQZ	CASSCF	2.0337	504	1.71
	SS-MRMPPT(RS)	2.0020	584	2.63
	SS-MRMPPT(BW)	1.9991	565	2.47
TZVP	LC-CEPA-3[a]	1.989	585.8	
	LC-CEPA-3+TQ/col/line[a]	2.016	533.9	
	LC-CEPA-3+TQ/explicit[a]	2.013	539.1	
	CCSD[a]	2.003	559.0	
	CCSD(T)[a]	2.011	543.1	
	experiment	1.988	564.9	2.475

[a] Ref 70. Experiment: ref 86 (2,2) CAS has been used in our works.

Table 7. Spectroscopic Constants for the Electronic Ground State of the Br_2 Molecule Using Different cc-pVXZ Bases[a]

basis	method	R_e (Å)	ω_e (cm^{-1})	D_e (eV)
cc-pVTZ	CASSCF	2.3408	294	1.48
	SS-MRMPPT(RS)	2.2854	326	2.33
	SS-MRMPPT(BW)	2.2892	328	2.18
cc-pVQZ	CASSCF	2.3371	294	1.50
	SS-MRMPPT(RS)	2.2684	336	2.47
	SS-MRMPPT(BW)	2.2739	334	2.25
cc-pV5Z	CASSCF	2.3371	293	1.50
	SS-MRMPPT(RS)	2.2720	331	2.47
	SS-MRMPPT(BW)	2.2786	327	2.20
	experiment	2.283	323.2	1.971

[a] Experiment: ref 86; (2,2) CAS has been used in our works.

calculations documented in Tables 6 and 7 have been done with the exclusion of the He core correlation for Cl_2 and Br_2. To judge the performance of the SS-MRMPPT methods for the Cl_2 system, we analyze our results vis-a-vis the values obtained from the CCSD, CCSD(T) LC-CEPA-3, and LC-CEPA-3+TQ methods.[70] Table 6 clearly demonstrates that our method yields equilibrium bond lengths and harmonic frequencies that are quite akin to those obtained from LC-CEPA-3 for Cl_2 molecule. Although LC-CEPA-3+TQ equilibrium bond lengths are similar to CCSD and CCSD(T), the average absolute deviation relative to experimental equilibrium bond lengths is a little larger than that of our results with the cc-pVQZ basis. We have obtained consistently quantitative descriptions for Cl_2 and Br_2 molecules along the wide range of bonding coordinates and thereby got encouraging spectroscopic constants with good accuracy. Comparing the experimental results, we see once again that the SS-MRMPPT method yields an accurate description of the spectroscopic parameters. All of the calculated bond lengths for Br_2 are slightly longer than the experimental result, and the reverse is true for Cl_2. For both the systems, the SS-MRMPPT calculations overshoot the vibrational frequency. We have observed that the spectroscopic constants for Cl_2 and Br_2 systems obtained from SS-MRMPPT with a good basis are chemically accurate. For both these diatomic systems, the performance of SS-MRMPPT(BW) is slightly better than the RS scheme as that in the case of the F_2. In

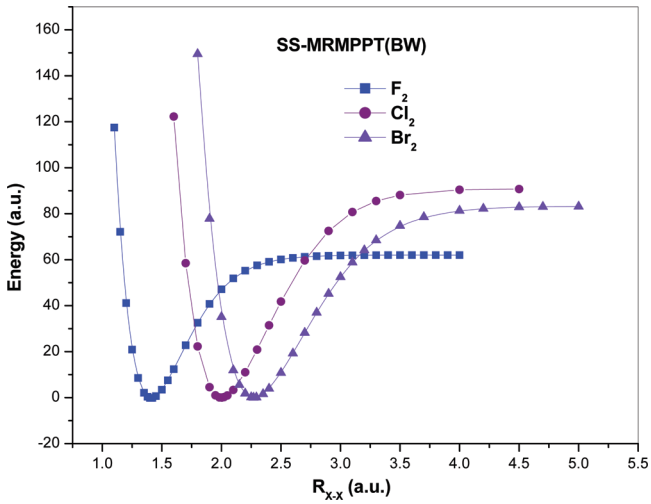


Figure 6. Potential energy surfaces for the X_2 [X = F, Cl, and Br] molecules with the SS-MRMPPT(BW) method in the cc-pVQZ basis set.

the more complex system, Br_2, we have a less satisfactory performance of the SS-MRMPPT method. While the calculated results shown here are not free from approximation, and thus not exact, it is nonetheless significant that, of all X_2, Br_2 shows the largest discrepancy between our calculation and the experimental data. The cause of this discrepancy is not clear. We ascribe the unsatisfactory result of the SS-MRMPPT for Br_2 (i) to the increased number of electrons [implying a rather large CAS] and (ii) to the absence of the relativistic effects in the formalism. Actually, in F_2, Cl_2, and Br_2, the active orbitals mainly consist of atomic p, d, and f orbitals, respectively. To increase the accuracy, one can suggest the inclusion of appropriate sets of orbitals in the active space. However, the resulting dimension of the CAS space would make the corresponding numerical calculations most unpractical, especially for Cl_2 and Br_2. Recent studies at the MRCI level[63] demonstrate the importance of the flexibility of the reference space in the description of PESs in the case of the F_2 molecule. We have already mentioned that the incorporation of the relativistic effect is very important for the numerical accuracy of the SS-MRMPPT method for large systems. This is also true for other nonrelativistic methods. For Cl_2 and Br_2 calculations, using a rather large CAS cannot even provide a qualitatively acceptable form of the PES due to the strong importance of the dynamical correlation of the inner electrons along with relativistic effects. In our opinion, in order to get a quantitative agreement of the spectroscopic constants with the corresponding experimental data, inclusion of these effects is inevitable. It is noteworthy that, in order to make contact with physical reality, using large basis sets and extrapolations to the complete basis set (CBS) limit is essential, though we have not explored this aspect in the current paper. A good deal of quantitative information has been learned about Cl_2 and Br_2 from this study, but a considerable amount of additional theoretical research should be performed in order to understand the molecular species better.

The calculated values of the dissociation energy exhibit the trend $Cl_2 > Br_2 > F_2$ [see also Figure 6], as also observed experimentally. This is usually explained by enhanced Pauli repulsion between the occupied p(π) orbitals, which is particularly strong in F_2 because it has the shortest bond of the dihalogens. The plots shown in Figure 6 need special mention at this juncture. Here, we have plotted the potential energy surfaces for the X2 system (X = F, Cl, Br), and for a clear presentation we have subtracted out the energy value at the equilibrium internuclear distance for the concerned system in an attempt to emphasize the actual trend in the data. It is quite evident that the computed equilibrium bond distances via SS-MRMMPPT(BW) and SS-MRMMPPT (RS) agree well with the experimental trend for the X2 systems, that is, $R_{eq}(F-F) > R_{eq}(Cl-Cl) > R_{eq}(Br-Br)$.

B. Ground State PES of BH Molecule. In this subsection, we consider the dissociation of the diatomic boron hydride (BH) molecule, where the presence of an open-shell model function(s) is necessary for the accurate description of nondynamical correlation. The accurate computation of the ground-state ($X^1\Sigma^+$) PES of the BH molecule is a "touchstone" for many ab initio methods.[18,52,71−81] The ground state reaction path of the BH system shows varying degrees of quasi-degeneracy with a rather physical nature and hence is appropriate to test the efficacy of the different multireference many-body methods. Thus, for this molecule, the use of the spin-free SS-MRPT method in studying PES seems to be justified. It is well documented that MP2, MP3, and MP4 schemes of the SR-based method cannot lead to a correct dissociation limit.[64] Kowalski and Piecuch[76] observed that the quality of the results of the CCSD[T], CCSD(T), and CCSD(TQ_f) methods is not as good at larger bond distances. They have demonstrated that the renormalized CC methods[82] are able to remove the failures of the traditional SRCC methods at a larger internuclear bond separation. Al-Saidi et al.[81] also observed the failure of RCCSD(T) with the cc-pVXZ basis to describe the BH molecule for larger bond lengths. They observed that AFQMC/cc-pVDZ results are in very good agreement with FCI energies and exhibit uniform behavior across the entire PES. Dutta and Sherrill[75] also showed that the performance of MP2 and CCSD(T) goes down with very large errors in the bond breaking region [the computed PES has an unphysical shape in the intermediate bond breaking region]. This arises from the inapplicability of SR-based theory in cases of strong degeneracies as occurring at large bond distances. Dutta and Sherrill[75] summarized that methods on the basis of the UHF reference exhibiting significantly better performance for BH molecule than the RHF reference. In our calculation, we have used the CASSCF reference, as it is able to dissociate BH in a correct manner.

To compute PES, for this molecule, we have employed two different basis sets, namely, 6- 31G** and cc-PVQZ, which enables a comparison with the exact FCI results.[52,80] In both cases, we employed a CAS(2,2) [(core)σ^2, (core)σ^{*2}, and (core)$\sigma\sigma^*$]. As already mentioned, we also present the calculations performed on the BH molecule employing the same CAS scheme [CAS(val)] and basis as used by Sherill and co-workers[52] for a comparison with the literature data of various previous calculations. The larger the space, the larger is the amount of the nondynamical correlation introduced. Calculations with a second active space, CAS-

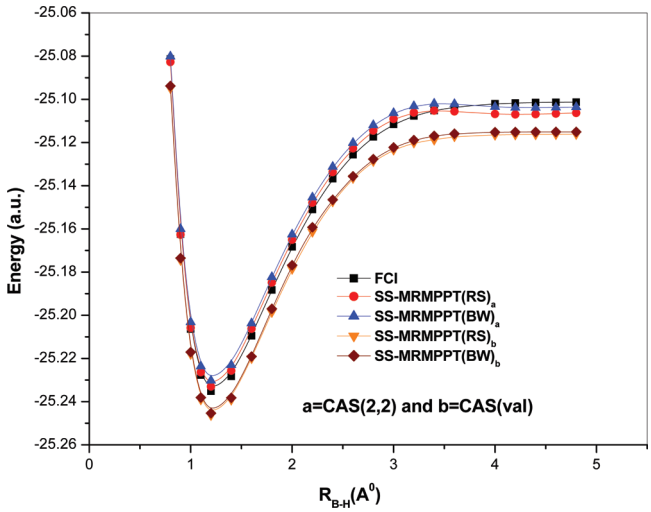


Figure 7. Potential energy surfaces for the BH molecule in the cc-pVQZ basis set.

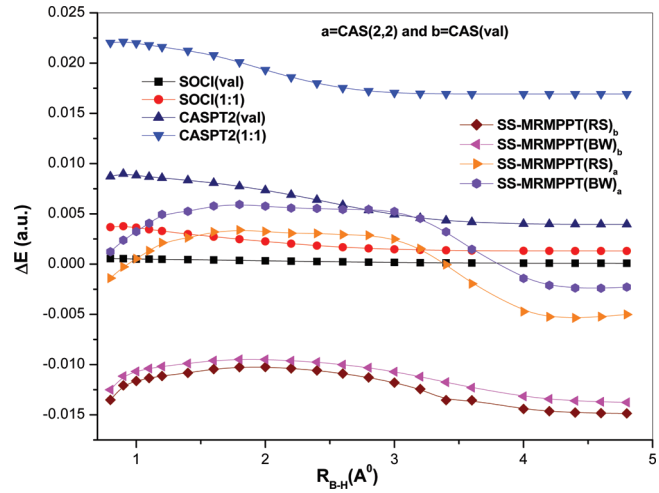


Figure 8. Plots of the errors versus FCI as a function of bond length using the cc-pVQZ basis set.

Table 8. Nonparallelity Error (NPE) in kcal/mol for BH Using Various Basis Sets[a]

basis	method	NPE
cc-pVQZ	CASSCF(val)	9.25
	SS-MRMPPT(RS)/CAS(val)	2.86
	SS-MRMPPT(BW)/CAS(val)	2.65
	SOCI(val)	0.29
	SOCI(1:1)	1.54
	CASPT2(val)	3.16
	CASPT2(1:1)	3.26
6-31G**	SS-MRMPPT(RS)/CAS(2,2)	4.82
	SS-MRMPPT(BW)/CAS(2,2)	4.76
	SS-MRMPPT(RS)/CAS(val)	2.89
	SS-MRMPPT(BW)/CAS(val)	2.68
	3R-BWPT2	7.38
	3R-BWPT2$_{corr}$	4.83
	BWCCSD	7.16

[a] SOCI and CASPT2 values have been taken from ref 52. BWPT and BWCC values have been taken from ref 80.

(val), have been performed in order to see whether this active space could give a more balanced treatment of the correlation effect for the target state. As CAS contains open-shell CSF, it is desirable to apply a proper spin-adapted state-specific theory to compute a completely accurate PES. The use of the cc-pVXZ basis set in our present calculations of spectroscopic constants is due to the possibility of direct comparison with calculations by other methods including the FCI one. Sherrill and co-workers[78,79] have provided several spectroscopic constants for the ground state of BH using different types of basis sets including the cc-pVXZ via FCI method.

Dissociation PESs for the BH molecule for the cc-pVQZ basis via the CASPT2, SOCI, and FCI methods have been reported by Abrams and Sherrill.[52] The PES obtained by the SS-MRMPPT method along with FCI are displayed in Figure 7. As far as the shape of the resulting PES is concerned, it is observed that the SS-MRMPPT/CAS(val) also produces a qualitatively correct PES, just as that of the FCI one. Therefore, we feel encouraged to investigate the PES using the larger reference space, CAS(val), and we are interested to see whether an increase of the reference space improves the results. Here, we want to mention that the quality of the SS-MRMPPT/CAS(val) PES is slightly better than the CASPT2(val) as far as NPE is concerned.

We now focus our attention to the error graph with respect to the corresponding FCI values. In Figure 8, we have plotted the corresponding graph along with previously published results using CASPT2 and SOCI methods.[52] In terms of these deviations, it has been observed that the CASPT2(1:1) performs much poorer than the methods considered here. The errors of SS-MRMPPT with respect to the FCI are modest, and overall SS-MRMPPT does quite well in the case of BH. Although SS-MRMPPT/CAS(2,2) values are more close to the FCI in comparison to the SS-MRMPPT/CAS(val), the errors for the former change sign in contrast to the latter one. In other words, a larger fluctuation for SS-MRMPPT/CAS(2,2) is observed relative to that for the SS-MRMPPT/CAS(val). Not only is the value of error an important issue, equally important is the requirement of the generated PES to be parallel to the corresponding FCI one. The deviation of SS-MRMPPT/CAS(val) from FCI is comparable with that of CASPT2(val). For better clarity, in Table 8, we also summarized the nonparallelity errors (NPE). For a given set of calculations in a dissociative surface, the NPE is defined as the difference between the maximal and minimal deviations from the exact FCI PES. The NPE for SS-MRMPPT approaches are very modest. In terms of the NPE (indicating the quality of the overall shape of the PES), we observe that the SS-MRMPPT/CAS(val) PES is marginally better than the CASPT2 method.

Recently, Paap et al.[80] have published applications of multireference state specific second-order Brillouin−Wigner perturbation theory (BWPT2) to the bond breaking process in the ground state of the BH molecule using the 6- 311G** basis. It is thus instructive to examine the performance of SS-MRMPPT(BW) using the same basis set. The results of the BWPT2 method are then used to access the performance of the calculations based on the SS-MRMPPT method. In Figures 9 and 10, we have summarized results for the 6-311G** basis as a function of the nuclear separation. Also, as expected, the RMP2[80] is seen to perform reasonably well at smaller internuclear separations but becomes increasingly

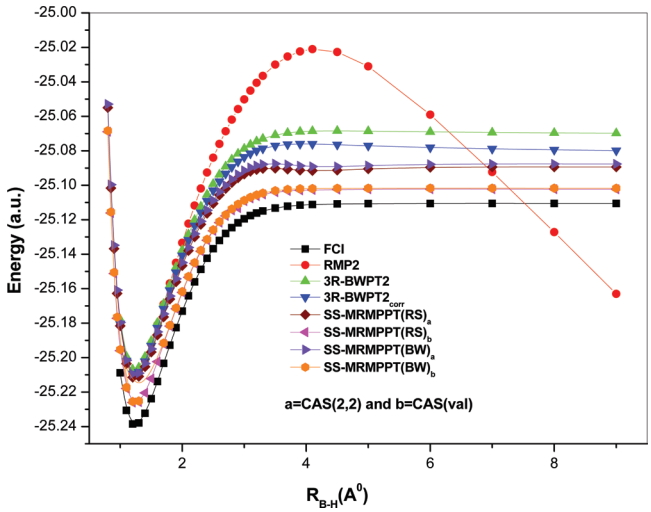


Figure 9. Potential energy surfaces for the BH molecule in the 6- 311G** basis set.

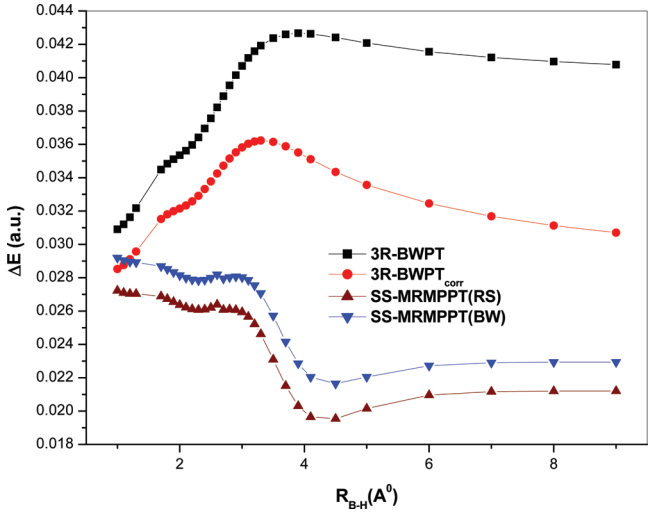


Figure 10. Plots of the errors versus FCI as a function of bond length using the 6-311G** basis set [CAS(2,2) has been used in SS-MRMPPT calculations].

poor (diverging in nature) as one goes over to large nuclear separations (owing to the poor quality of the RHF wave function in describing bond breaking), indicating the necessity of a multireference description. Figure 10 illustrates that the BWPT2 energies are in less agreement with FCI than the SS-MRMPPT with CAS(2,2) and CAS(val). In the 6-311G** basis, SS-MRMPPT with CAS(val) gives a better overall accuracy and a more uniform behavior than SS-MRMPPT/CAS(2,2) in mapping the PES. This is also evident from the NPE values. The NPE values calculated for various methods using the 6- 311G** basis are set out in Table 8. Inspecting the results listed in Figures 9 and 10 as well as the NPE values in Table 8, we can draw the conclusion that the PESs have been generated reasonably well by the SS-MRMPPT method. Considering the overall performance (in the context of NPE nd spectroscopic constants) in the entire studied region of geometries for various basis sets, the SS-MRMPPT/CAS(val) scheme gives very satisfactory results with the small NPE.

In our numerical work, we thus observed that SS-MRMPPT produces a smooth and consistent behavior across the entire PES between the equilibrium and the dissociation limit in BH bond breaking for various cc-pVXZ bases, suggesting it may be used with confidence to calculate various spectroscopic constants. In Table 9, the SS-MRMPPT results are gathered and compared with those computed in the most recent theoretical studies; the experimental data are also reported. Using recent theoretical results, we have summarized spectroscopic constants calculated by CCSD(T) and FCI (acts as benchmark) values for the various cc-pVXZ basis sets of Sherrill and co-workers[79] in the same table. We have also included the results for various basis sets employing different methods considered here which can provide a measure of the effects due to basis set choice (i.e., which level of basis set is required to obtain a given accuracy). The choice of an appropriate basis set is especially intricate in studies on electronic states of a molecule. In order to examine the accuracy of the spectroscopic values at different levels of correlated theory, we also tabulated the coupled cluster results of Martin et al.[73] and Larsen et al.[77] with different basis sets to display the effectiveness of the SS-MRMPPT method, as a scheme, to provide spectroscopic properties. They have demonstrated the convergence aspect of spectroscopic constants of BH with respect to contracted and uncontracted basis sets. They have also illustrated that the correction due to the nonadiabatic effect is much more important in predicting spectroscopic constants accurately than that of basis set errors in the case of BH. In this context, we mention the recent work of Sherrill and coworkers.[78,79]

From the tabulation of data in Table 9, we observed that for each basis set the performance of the SS-MRMPPT methods in both the perturbative schemes is very close to the results of CCSD(T) and FCI for Sherrill and co-workers.[78,79] Here, we should recall that the SS-MRMPPT method is computationally quite less demanding compared to the significant computational cost of the CC-based methods. The works of Abrams and Sherrill[78] via the FCI/DZP-NO(5Z) scheme provide very sound estimations of the ground state dissociation energy, vibrational frequency, equilibrium internuclear distance, and some other measurable quantities. As we have seen for the BH molecular system, SS-MRMPPT can also compete with the precision of the other well established sophisticated *ab initio* methods. At this point, we want to discuss the accuracy of the SS-MRMPPT method to yield dissociation energy, D_e. Our best theoretical value for D_e is 3.57 eV, obtained by SS-MRMPPT(RS)/cc-pVTZ. Bauschlicher et al.[74] determined a D_e of 3.65 eV using (4e-/9) orbital complete active space multireference configuration interaction wave functions. A more accurate D_e value was computed by Curtiss and Pople[72] through QCISD(T) calculations. The most recent calculation by Miliordos and Mavridis[83] provides a value for D_e of 3.53 eV. Our works thus provide very sound estimations of the ground state dissociation energy of BH. As far as the computational cost is concerned, the overall agreement of the SS-MRMPPT D_e with accurate high-level theoretical methods is very good, giving general support to the reliability of the present SS-MRMPPT results. From a comparative study of various

Table 9. Spectroscopic Constants for the Electronic Ground State of BH Molecule Using Different cc-pVXZ Bases[a]

basis	method	R_e (Å)	ω_e (cm^{-1})	D_e (eV)
cc-pVDZ	CASSCF	1.2672 (1.2665)	2269 (2277)	3.25 (3.12)
	SS-MRMPPT(RS)	1.2582 (1.2552)	2310 (2347)	3.35 (3.28)
	SS-MRMPPT(BW)	1.2581 (1.2554)	2310 (2346)	3.34 (3.28)
	CCSD(T)[b]	1.2558	2342.65	
	FCI[b]	1.2560	2340.72	3.44
cc-pVTZ	CASSCF	1.2512 (1.2504)	2275 (2278)	3.36 (3.18)
	SS-MRMPPT(RS)	1.2296 (1.2267)	2348 (2372)	3.54 (3.48)
	SS-MRMPPT(BW)	1.2299 (1.2273)	2355 (2378)	3.54 (3.45)
	CCSD(T)[b]	1.2354	2350.84	
	FCI[b]	1.2356	2348.71	
cc-pVQZ	CASSCF	1.2498 (1.2242)	2270 (2392)	3.37 (3.06)
	SS-MRMPPT(RS)	1.2317 (1.2227)	2349 (2395)	3.55 (3.46)
	SS-MRMPPT(BW)	1.2314 (1.2222)	2354 (2377)	3.55 (3.46)
	CCSD(T)[b]	1.2333	2358.91	
	FCI[b]	1.2335	2356.78	
	FCI/cc-pV 5Z[b]	1.23285	2358.21	
	FCI/cc-pCVDZ[b]	1.25434	2340.12	
	FCI/cc-pCVTZ[b]	1.23339	2355.26	
	FCI/cc-pCVDZ[c]		2340.1	
	FCI/aug-cc-pCVDZ[c]		2320.8	
	MRCI[d]	1.2301	2358	3.68
	MRCI+Q[d]	1.2301	2359	3.68
	RCCSD(T)[d]	1.2296	2361	3.68
	RCCSDT[d]	1.2304		3.68
	FCI/DZP[e]	1.2491	2339	3.48
	FCI/DZP-NO[e]	1.2362	2354	3.57
	FCI/DZP-NO(5Z)[e]	1.2362	2350	3.57
	FCI/631G**[e]	1.2344	2388	3.61
	FCI/WMR-ANO[e]	1.2675	2309	3.47
	ref 71	1.2338	2254	3.65
	exp (ref 87)	1.23217	2366.73	
	experiment	1.2324	2367	3.57

[a] The values in parentheses describe the values using CAS(2,2). [b] Ref 79. [c] Ref 73. [d] Ref 83. [e] Ref 78, experiment: ref 86.

spectroscopic properties, we have found that SS-MRMPPT is a reliable tool to achieve quantitative accuracy for calculating various spectroscopic properties of BH via the computation of intermolecular interaction energies over all geometries including those near the dissociation limit, as with that of X_2 systems.

We now discuss the aspect of the effect of orbital rotation on the SS-MRMPPT energy from a numerical point of view for the BH system. Whether or not a method is invariant depends on the ansatz and the amplitude equations. It is now well documented that there are serious difficulties in making JM-ansatz based methods orbital invariant with respect to active space rotation, although they are invariant with respect to rotations in core orbital space and inactive virtual orbital space.[41,84] This is a very relevant and important issue. If a molecule has degenerate representations at a high symmetry point, then it may happen very easily that the orbitals change very rapidly as a function of minor geometrical distortions. If the electronic structure method is sensitive to such rotations (for example, if the results depend on the precise definitions of the orbitals at the high symmetry point), then one can expect to obtain poor results. In our numerical analysis, we have observed that the change in SS-MRMPPT energy due to the active−active rotation is not negligible, and the change for SS-MRMPPT(BW) is less than its RS counterpart [see Table 10]. Our results indicate that the change in energy due to virtual−virtual rotations is much smaller than that of the active−active one. On the other hand, the effect of occupied−occupied orbital(s) rotations exhibits a very small influence. Thus, the resulting effective Hamiltonians, $\tilde{H}^{(2)}_{\mu\nu}$, of SS-MRMPPT method 5 are not invariant with respect to orbital rotations within active subspace, and therefore, to ensure reproducibility of the energies, the orbitals should be specified unambiguously.

Table 10. Numerical Test of the Change in Energy (a.u.) of SS-MRMPPT/CAS(2,2) with Respect to the Orbital Rotations (about $\pi/4$), Performed on the Ground State of the BH Molecule for the Equilibrium Bond Length in the cc-pVQZ Basis[a]

rotation	SS-MRMPPT(RS)	SS-MRMPPT(BW)
no rotation	−25.214396	−25.212309
A−A rotation	−25.220023	−25.214260
V−V rotation	−25.214455	−25.212350

[a] A−A, active−active; V−V, virtual−virtual.

Our study explores several general trends of MR-based methods. Figures of PESs show that the SS-MRPT procedure is able to yield a qualitatively balanced description for the equilibrium region as well as the dissociation zone of the X_2 and BH molecule, and the computed dissociation PESs are completely exempt from any "intruder state" problem. Our work in this paper shows that the performance of the SS-MRMPPT(BW/RS) methods to compute the spectroscopic constants of different chemical systems with varying degrees of MR character is quite satisfactory. Applications to a number of state energies and comparison with benchmark FCI results (when available) show a uniform behavior of the SS-MRMPPT formalism. We see that SS-MRMPPT recovers most of the binding energy of the X_2 and BH

molecules. This fact is more significant if one considers the dimension of CAS we have used. The agreement of the SS-MRMPPT spectroscopic constants with the previously published theoretical calculations is reasonably good. The accuracy of the results appears to be encouraging, taking into account the low computational cost. We have illustrated that the method gives geometries, frequency, and dissociation energies which are at least as accurate as the corresponding results obtained with the Mk-MRCC method and more reliable than the CCSDT method for X_2 systems where the wave function has a strong MR character in their ground state.

From the foregoing analysis we have noticed that the two methods, SS-MRMPPT(RS) and SS-MRMPPT(BW), are also chemically accurate relative to each other. However, the performance of SS-MRMPPT(BW) is generally better than that of its RS cousin, although the RS scheme is computationally less expensive than its BW counterpart. Further work is required to assess more fully the accuracy of the SS-MRMPPT(BW) method; however, our initial results validate the accuracy of the theory vis-à-vis the experiment and other theoretical methods. Very good agreement with the experiment is observed for F_2, Cl_2, and BH molecules using the SS-MRMPPT technique, while a relatively less satisfactory agreement is observed for Br_2. As a whole, the above numerical analysis reveals that the SS-MRMPPT is a useful method, with an extremely reasonable performance/cost ratio.

We now discuss the effect of basis set size. As is well-known, the choice of basis used for calculations of spectroscopic constants is very crucial, especially in the determination of bond dissociation energy. A closer inspection of the comparison of the computed spectroscopic constants for the various basis sets gives us an optimistic view of the expected accuracy of the individual orbital basis sets and of the infinite basis set extrapolation limit. The spectroscopic constants summarized in the tables display the necessity and utility of using large basis sets in analyzing and demonstrating intrinsic errors associated with theoretical methods. The SS-MRMPPT calculations have been performed for X_2 with different basis sets using the same active space to examine the basis set effect on the quality of spectroscopic constants. The systematic exploration of various basis sets permits an assessment of the reliability of our results. Although the shape of the surface is similar for all the basis sets, there are important quantitative changes of PESs as the basis set is increased [see Figures 1 and 2]. From the tables, one notes that the variation of basis sets has a substantial effect on spectroscopic constants (which looks rather unsystematic at first glance), as it should be. From the tables of F_2, the change of spectroscopic constants appears slightly less systematic when compared to the Mk-MRCC approach. This is also true for other systems. We hope to investigate the issue more deeply in the future. However, the aforementioned findings encourage us to make the comment that the numerical accuracy of the SS-MRMPPT method is appreciably good, and the method is capable of incorporating most of the essential correlation that has been left out by the mean-field method(s). It should be noted that our calculations have not been corrected for basis-set superposition errors (BSSE) point-by-point using the counterpoise technique. Such a correction reduces an artificial bias toward dissociation energies and bond lengths, particularly for weakly bound species, such as those investigated in this work. It is worth noting, in the dissociation PES calculations of F_2, that the contributions due to innercore electron (1s) correlation and relativistic effects (which have been excluded from the correlation treatment in the present study) to the dissociation of the molecule have been estimated to be less than 1 kcal/mol.[26,85] The basis set superposition error in the case of the F_2 molecule is about the same in magnitude but opposite in sign. So one can expect the cancellation of these effects for the computation of the dissociation PES of F_2 [see ref 85]. Nevertheless, incorporation of relativistic effects to chemical processes is not only desirable but also essential for the level of "spectroscopic accuracy" or "chemical accuracy" even for X_2 and BH diatomics. Considering the pros and cons of the SS-MRPT approach, we think that further improvements are needed. However, it should also be pointed out that the corrections due to the core correlation, nonadiabatic correction, basis saturations, and relativistic(s) effects are not always necessarily negligible compared to the intrinsic errors in the methods considered.

IV. Conclusion

This paper is a continuation of our preceding studies[40] on the numerical applications of the SS-MRPT method, drawing on a previous work of Mukherjee and co-workers.[14] The SS-MRPT method, which is based on a multiconfiguration reference state, can provide increased accuracy for treating potential energy surfaces far from equilibrium, certain types of excited states, and the mapping of complete reaction paths. The SS-MRMPPT method includes nondynamical and dynamical correlation effects in a balanced way in the electronic wave function of closed-shell and open-shell states and performs satisfactorily at low numerical expenses (the only problem seems to be a rapidly increasing cost of CASSCF with the increasing size of the active space). For large systems, SS-MRPT is usually the method of choice instead of SS-MRCC (can only be applied to relatively small electron systems because of their high computational cost) as the former gives a manageable accuracy/cost ratio for dealing with multiconfigurational problems.

A characteristic application of SS-MRPT is to describe bond dissociation processes in rather complicated cases with a satisfactory accuracy. The SS-MRMPPT in RS and BW variants is endowed with the desirable properties of strict separability and *absence of intruder states*. The SS-MRPT can handle the lowest energy state regardless of charge or spin symmetry with reasonable and consistent accuracy, supporting our use of this method as a "standard" for treating small- to medium-sized systems. As the validity of the SS-MRMPPT(RS) has already been illustrated in a large number of earlier publications,[40] this work focuses on establishing the efficacy of the SS-MRMPPT(BW) method. We have included the results of SS-MRPT, using Møller−Plesset (MP) as well as Epstein−Nesbet (EN) partitions with respect to

the RS and BW expansion. In order to show that our results are not spoiled by intruder state effects that are not related to the partitioning, intruder containing molecules and reference spaces have been investigated here.

In this work, we performed SS-MRMPPT calculations using complete active space: keeping the active space as small as possible for computing the nondynamical correlation (i.e., proper dissociation). The aim of the present investigation is to calibrate the adequacy of the SS-MRPT approach through the computation of the reaction paths of singlet ground states of X_2 [X = F, Cl, and Br] and BH for Dunning's correlation-consistent double-, triple-, and quadruple-ζ basis sets. As recognized in several studies, along the dissociating reaction path of X_2 and BH, the zeroth-order reference function changes multiconfigurationally, and hence the bond breaking involved is difficult to handle computationally. Therefore, the dissociation of X_2 and BH molecular systems is a demanding test case used traditionally to benchmark new computational methods. In order to display the performance of the SS-MRMPPT method, the potential energy surface obtained using the SS-MRPT method (without changing the size of the CAS) is used to calculate the spectroscopic constants. The accuracy of the computed SS-MRMPPT spectroscopic constants is assessed by comparing them with the corresponding accurate and established theoretical and experimental results (whenever available). In general, agreement between the SS-MRMPPT and experiment is significantly better than that with its EN counterpart for the systems studied by us in this paper. With the example applications presented here, it seems that it may not be fair at this stage to conclude definitively about the relative performance of SS-MRMPPT and SS-MRENPT methods. More exhaustive calculations, in particular for the SS-MRENPT, are needed to come to a definitive conclusion, which is on the way. In the present work, the applicability of the SS-MRMPPT(BW) method to X_2 and BH systems is documented in detail. In the case of F_2 system, we consider results from full-blown Mk-MRCC (parent SS-MRCC theory) and BW-MRCC calculations for a comparison reported recently which establishes the fact that the SS-MRMPPT method provides a convenient way (considering the rather low computational effort) to generate an accurate potential energy surface involving bond breaking to provide spectroscopic constants of good quality for the ground state of F_2. It has been demonstrated that the SS-MRMPPT(BW) along with its RS counterpart provide very good results for single bond breaking over the entire reaction pathways, eliminating the failures of the conventional MRPT methods in those multireference situations. It is to be remarked that our SS-MRMPPT results for the RS and BW expansion produce spectroscopic properties close to each other even in the smallest CAS space. The foregoing numerical analysis of the SS-MRMPPT(RS) and SS-MRMPPT(BW) results demonstrates clearly that the BW values are better as compared to the RS values, as is seen in the NPE value. The different spectroscopic constants that we have computed too point to this. Our numerical results also confirm that the orbital rotation effect on the energy is less in the case of SS-MRMPPT(BW) in comparison to its RS counterpart. Thus, the SS-MRMPPT(BW) method can be considered as a very effective perturbative companion of the state-specific multireference coupled cluster method of Mukherjee and coworkers.

The issue that merits separate discussion at this juncture is the extent of relaxation of the reference coefficients in SS-MRMPPT. The effect of large mixing of reference functions and consequent relaxation cannot be estimated fully by SS-MRMPPT(RS), while the SS-MRPT(BW) scheme takes this relaxation fully during the computation of cluster amplitudes and energy. Here it should be mentioned that, if the computational model does not allow relaxation of the coefficients of the reference configurations in the correlation treatment, a correct description of the potential energy surfaces where the orbitals change very rapidly as a function of minor geometrical distortions (as in the case of weakly avoided crossings) cannot be expected. We plan to explore this issue in the near future by considering systems that pose such complexities.

A source of error in our SS-MRPT results (especially in the context of Br_2) is the exclusion of the relativistic effects. Additionally, a proper and good description of the electronic structure of the X_2 must involve an improvement of construction of the zero-order wave function. Our findings demonstrate that the SS-MRPT approach can indeed accommodate most of the effects responsible for binding in the F_2, Cl_2, Br_2, and BH molecules. This paper presents an effort toward the ongoing research to produce the accurate spectroscopic parameters for F_2, Cl_2, Br_2, and BH which one can use in a variety of spectroscopic and chemical applications. The present results should prove useful in the calibration of new theoretical methods for bond breaking. In order to achieve chemical accuracy for our computed spectroscopic constants with respect to the experimental findings, incorporation of relativistic effects is an inevitable issue. Our observations suggest that the SS-MRMPPT is a very reasonable and useful variant of MRPT with comparable strengths, and the corresponding wave function has sufficient flexibility to model the large changes in electronic structure that can occur during chemical reactions. As a final remark, we can say that the success of the method ensures that a great deal of additional work on the SS-MRPT programs is to be expected in the days to come.

Acknowledgment. We would like to thank the anonymous reviewer for critical reading of our paper and various crical suggestions. We gratefully acknowledge Department of Science and Technology (India) funding for this work [Grant No.SR/S1/PC-32/2005].

References

(1) Sheppard, M. G.; Freed, K. F.; Herman, M. F.; Yeager, D. L. *Chem. Phys. Lett.* **1979**, *61*, 577. Freed, K. F. In *Lecture Notes in Chemistry*; Kaldor, U., Ed.; Springer: Berlin, 1989; Vol. 52, p 1.

(2) Nitzsche, L. E.; Davidson, E. R. *J. Chem. Phys.* **1978**, *68*, 3103. Nitzsche, L. E.; Davidson, E. R *J. Am. Chem. Soc.* **1978**, *100*, 7201. Rawlings, D. C.; Davidson, E. R. *Chem. Phys. Lett.* **1983**, *98*, 424. Davidson, E. R.; McMurchie, L. E.; Day, S. J. *J. Chem. Phys.* **1981**, *74*, 5491. Davidson, E. R.

J. Chem. Phys. **1968**, *48*, 3169. Murray, C. W.; Davidson, E. R. *Chem. Phys. Lett.* **1991**, *187*, 451.

(3) Hirao, K. *Chem. Phys. Lett.* **1992**, *190*, 374. Hirao, K. *Chem. Phys. Lett.* **1992**, *196*, 397. Hirao, K. *Int. J. Quantum Chem.* **1992**, *S26*, 517. (a) Hirao, K.; Nakano, H.; Hashimoto, T. *Chem. Phys. Lett.* **1995**, *235*, 430. (b) Hashimoto, T.; Nakano, H.; Hirao, K. *J. Mol. Struct.: THEOCHEM* **1998**, *451*, 25. Choe, Y.-K.; Nakao, Y.; Hirao, H. *J. Chem. Phys.* **2001**, *115*, 621. Nakao, Y.; Choe, Y.-K.; Nakayama, K.; Hirao, K. *Mol. Phys.* **2002**, *100*, 729.

(4) Nakano, H. *J. Chem. Phys.* **1993**, *99*, 7983.

(5) Andersson, K.; Malmqvist, P. Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Chem. Phys.* **1990**, *94*, 5483.

(6) Werner, H. *Mol. Phys.* **1996**, *89*, 645.

(7) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3647.

(8) Celani, P.; Stoll, H.; Werner, H.-J.; Knowles, P. J. *Mol. Phys.* **2004**, *102*, 2369.

(9) Finley, J. P.; Chaudhuri, R. K.; Freed, K. F. *Phys. Rev.A* **1996**, *54*, 343. Zaitevskii, A.; Malrieu, J. P. *Theor. Chim. Acta* **1997**, *96*, 269.

(10) Malrieu, J.-P.; Heully, J.-L.; Zaitevskii, A. *Theor. Chim. Acta* **1995**, *90*, 167.

(11) Angeli, C.; Cimiraglia, R.; Evangelisti, S.; Leininger, T.; Malrieu, J. P. *J. Chem. Phys.* **2001**, *114*, 10252. Angeli, C.; Cimiraglia, R.; Malrieu, J.-P. *J. Chem. Phys.* **2002**, *117*, 9138.

(12) Hubač, I.; Mach, P.; Wilson, S. *Mol. Phys.* **2002**, *100*, 859. Hubač, I.; Mach, P.; Papp, P.; Wilson, S. *Mol. Phys.* **2004**, *102*, 701. Papp, P.; Mach, P.; Pittner, J.; Hubač, I.; Wilson, S. *Mol. Phys.* **2006**, *104*, 2367. Papp, P.; Mach, P.; Hubač, I.; Wilson, S. *Int. J. Quantum Chem.* **2007**, *107*, 2622. Papp, P.; Neogrády, P.; Mach, P.; Pittner, J.; Hubacč, I.; Wilson, S. *Mol. Phys.* **2008**, *57*, 106.

(13) Wenzel, W.; Steiner, M. M.; Wilkins, J. W.; Wilson, K. G. *Int. J. Quantum Chem.* **1996**, *S30*, 1325. Wenzel Steiner, W. *J. Chem. Phys.* **1998**, *108*, 4714.

(14) Mahapatra, U. S.; Datta, B.; Mukherjee, D. *J. Phys. Chem.* **1999**, *103*, 1822.

(15) Ghosh, P.; Chattopadhyay, S.; Jana, D.; Mukherjee, D. *Int. J. Mol. Sci.* **2002**, *3*, 733.

(16) Ten-no, S. *J. Phys. Chem.* **2007**, *126*, 014108.

(17) Hoffmann, M. R. *J. Phys. Chem.* **1996**, *100*, 6125. Jiang, W.; Khait, Y. G.; Hoffmann, M. R *J. Phys. Chem. A* **2009**, *113*, 4374.

(18) Pahari, D.; Chattopadhyay, S.; Das, S.; Mukherjee, D.; Mahapatra, U. S. In *Theory and Applications of Computational Chemistry: The First 40 Years*; Dytkstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E.; Eds.; Elsevier: Amsterdam, 2005; p 581.

(19) Kozlowski, P. M.; Davidson, E. R. *J. Chem. Phys.* **1994**, *100*, 3672.

(20) Jarzecki, A. A.; Davidson, E. R. *J. Phys. Chem. A* **1998**, *102*, 4742.

(21) Leininger, M. L.; Allen, W. D.; Schaefer, H. F., III; Sherrill, C. D. *J. Chem. Phys.* **2000**, *112*, 9213.

(22) Olsen, J.; Fülscher, M. P. *Chem. Phys. Lett.* **2000**, *326*, 225.

(23) Zarrabian, S.; Laidig, W. D.; Bartlett, R. J. *Phys. Rev. A* **1990**, *41*, 4711.

(24) Rintelman, J. M.; Adamovic, I.; Varganov, S.; Gordon, M. S. *J. Chem. Phys.* **2005**, *122*, 044105.

(25) Azizi, Z.; Roos, B. O.; Veryazova, V. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2727.

(26) Pittner, J.; Šmydke, J.; Čársky, P.; Hubać, I. *J. Mol. Struct.: THEOCHEM* **2001**, *547*, 239.

(27) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3647.

(28) Murphy, R. B.; Messmer, R. P. *J. Chem. Phys.* **1992**, *97*, 4170.

(29) Dyall, K. G. *J. Chem. Phys.* **1995**, *102*, 4909.

(30) Robinson, D.; McDouall, J. J. W. *Mol. Phys.* **2006**, *104*, 681.

(31) Rolik, Z.; Szabados, Á.; Surján, P. R. *J. Chem. Phys.* **2003**, *119*, 1922. Surján, P. R.; Rolik, Z.; Szabados, Á.; Köhalmi, D *Ann. Phys. (Leipzig)* **2004**, *13*, 223.

(32) Rosta, E.; Surján, P. R. *J. Chem. Phys.* **2002**, *116*, 878.

(33) Grimme, S.; Waletzke, M. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2075.

(34) Choe, Y. K.; Nakao, Y.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 621.

(35) Chaudhuri, R. K.; Freed, K. F.; Chattopadhyay, S.; Mahapatra, U. S. *J. Chem. Phys.* **2008**, *128*, 144304.

(36) Evangelisti, S.; Daudey, J. P.; Malrieu, J.-P. *Chem. Phys.* **1983**, *75*, 91. Cimiraglia, R.; Persico, M. *J. Comput. Chem.* **1987**, *8*, 39.

(37) Jeziorski, B.; Monkhorst, H. *J. Phys. Rev.* **1981**, *24*, 1668.

(38) Zaitsevskii, A.; Malrieu, J. P. *Chem. Phys. Lett.* **1995**, *233*, 597. Zaitsevskii, A.; Malrieu, J. P. *Chem. Phys. Lett.* **1996**, *250*, 366.

(39) Mahapatra, U. S.; Datta, B.; Mukherjee, D. *J. Chem. Phys.* **1999**, *110*, 6171.

(40) Mahapatra, U. S.; Chattopadhyay, S.; Chaudhuri, R. K *J. Chem. Phys.* **2008**, *129*, 024108. Mahapatra, U. S.; Chattopadhyay, S.; Chaudhuri, R. K *J. Chem. Phys.* **2009**, *13*, 014101.

(41) Evangelista, F. A.; Simmonett, A. C.; Schaefer, H. F., III; Mukherjee, D.; Allen, W. D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4728.

(42) Angelia, C.; Calzado, C. J.; Cimiraglia, R.; Malrieu, J.-P. *J. Chem. Phys.* **2006**, *124*, 234109.

(43) Staroverov, V. N.; Davidson, E. R. *Chem. Phys. Lett.* **1998**, *296*, 435.

(44) Barbosa, A. G. H.; Barcelos, M. *Theor. Chem. Acc.* **2009**, *122*, 51.

(45) Kowalski, K.; Piecuch, P. *Chem. Phys. Lett.* **2001**, *344*, 165. Musia, M.; Bartlett, R. J. *J. Chem. Phys.* **2005**, *122*, 224102.

(46) Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571. Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045; see www.emsl.pnl.gov/ forms/basisform.html.

(47) Piecuch, P.; Paldus, J. *Theor Chim. Acta* **1992**, *83*, 69. Piecuch, P.; Paldus, P *J. Chem. Phys.* **1994**, *101*, 5875. Jeziorski, B.; Paldus, J.; Jankowski, P. *Int. J. Quantum Chem.* **1995**, *56*, 29.

(48) Evangelista, F. A.; Allen, W. D.; Schaefer, H. F., III. *J. Chem. Phys.* **2007**, *127*, 024102.

(49) Hanrath, M. *J. Chem. Phys.* **2005**, *123*, 84102. Engels-Putzka, A.; Hanrath, M. *Mol. Phys.* **2009**, *107*, 143.

(50) Máik, J.; Hubać, I. In *Quantum Systems in Chemistry and Physics: Trends in Methods and Applications*; McWeeny,

R., Maruani, J., Smeyers, Y. G., Wilson, S., Eds.; Kluwer Academic: Dordrecht, The Netherlands, 1997; pp 283. Pittner, J.; Demel, O.; Čársky, P.; Hubać, I. *Int. J. Quantum Chem.* **2002**, *90*, 1031.

(51) Pittner, J.; Gonzalez, H. V.; Gdanitz, R. J.; Čársky, P. *Chem. Phys. Lett.* **2004**, *386*, 211.

(52) Abrams, M. L.; Sherrill, C. D. *J. Phys. Chem. A* **2003**, *107*, 5611.

(53) Davidson, E. R.; Bender, C. F. *Chem. Phys. Lett.* **1978**, *59*, 369.

(54) Laidig, W. D.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 887.

(55) Nooijen, M. *Int. J. Mol. Sci.* **2002**, *3*, 656.

(56) Li, X.; Paldus, J. *J. Chem. Phys.* **1998**, *108*, 637. Li, X.; Paldus, J. *J. Chem. Phys.* **2006**, *125*, 164107.

(57) Mášik, J.; Hubač, I.; Mach, P. *Int. J. Quantum Chem.* **1995**, *53*, 297. Mášik, J.; Hubač, I.; Mach., P. *J. Chem. Phys.* **1998**, *108*, 6571.

(58) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *350*, 522. Krylov, A. I.; Sherrill, C. D. *J. Chem. Phys.* **2002**, *116*, 3194. Sears, J. S.; Sherrill, C. D.; Krylov, A. I. *J. Chem. Phys.* **2003**, *118*, 9084.

(59) Rosta, E.; Surján, P. R. *J. Chem. Phys.* **2002**, *116*, 878.

(60) Purwanto, W.; Al-Saidi, W. A.; Krakauer, H.; Zhang, S. *J. Chem. Phys.* **2008**, *128*, 114309.

(61) Yang, J.; Hao, Y.; Li, J.; Zhou, C.; Mo, Y. *J. Chem. Phys.* **2005**, *122*, 134308.

(62) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2005**, *122*, 224102. Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2006**, *125*, 204105.

(63) Bytautas, L.; Nagata, T.; Gordon, M. S.; Ruedenberg, K. *J. Chem. Phys.* **2007**, *127*, 164317. Bytautas, L.; Matsunaga, N.; Nagata, T.; Gordon, M. S.; Ruedenberg, K. *J. Chem. Phys.* **2007**, *127*, 204313.

(64) Shepard, R.; Kedziora, G. S.; Lischka, H.; Shavitt, I.; Müller, T.; Szalay, P. G.; Kállay, M.; Seth, M. *Chem. Phys.* **2008**, *349*, 37.

(65) Malrieu, J. P.; Spiegelmann, F. *Theor. Chem. Acta* **1979**, *52*, 55. Angeli, C.; Cimiraglia, R. *Theor. Chem. Acc.* **2002**, *107*, 313.

(66) Angeli, C.; Cimiraglia, R.; Malrieu, J.-P. *Chem. Phys. Lett.* **2000**, *317*, 472.

(67) Louderaj, U.; Harbola, M. K.; Sathyamurthy, N. *Chem. Phys. Lett.* **2002**, *366*, 88.

(68) Polyansky, O. L.; Császár, A. G.; Shirin, S. V.; Zobov, N. F.; Barletta, P.; Tennyson, J.; Schwenke, D. W.; Knowles, P. J. *Science* **2003**, *299*, 539. Cardoen, W.; Gdanitz, R. J. *J. Chem. Phys.* **2005**, *123*, 024304. Ruden, T. A.; Helgaker, T.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2004**, *121*, 5874.

(69) Sanchez-Marin, J.; Nebot-Gil, I.; Maynau, D.; Malrieu, J.-P. *Theor. Chim. Acta* **1995**, *92*, 241.

(70) Zhang, H.; Malrieu, J.-P.; Reinhardt, P.; Ma, J. *J. Chem. Phys.* **2010**, *132*, 034108.

(71) Gagliardi, L.; Bendazzoli, G. L.; Evangelista, S. *Mol. Phys.* **1997**, *91*, 861.

(72) Curtiss, L. A.; Pople, J. A. *J. Phys. Chem.* **1989**, *90*, 2522.

(73) Martin, J. M. L.; FranCois, J. P.; Gijbels, R. *J. Chem. Phys.* **1989**, *91*, 4425. Martin, J. M. L *Chem. Phys. Lett.* **1998**, *283*, 283.

(74) Bauschlicher, C. W., Jr.; Langhoff, S. R.; Taylor, P. R. *J. Phys. Chem.* **1990**, *93*, 502.

(75) Krylov, A. I.; Sherrill, C. D.; Byrd, E. F. C.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 10669. Dutta, A.; Sherrill, C. D. *J. Chem. Phys.* **2003**, *118*, 1610.

(76) Kowalaski, K.; Piecuch, P. *Chem. Phys. Lett.* **2001**, *344*, 165.

(77) Koch, H.; Christiansen, O.; Jørgensen, P.; Olsen, J. *Chem. Phys. Lett.* **1995**, *244*, 75. Larsen, H.; Olsen, J.; Jørgensen, P.; Gauss, J *Chem. Phys. Lett.* **2001**, *342*, 200.244.75. .

(78) Abrams, M. L.; Sherrill, C. D. *J. Chem. Phys.* **2003**, *118*, 1604.

(79) Temelso, B.; Valeev, E. F.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 3068.

(80) Papp, P.; Mach, P.; Pittner, J.; Hubač, I.; Wilson, S. *Mol. Phys.* **2006**, *104*, 2367.

(81) Al-Saidi, W. A.; Zhang, S.; Krakauer, H. *J. Chem. Phys.* **2007**, *127*, 144101.

(82) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2000**, *113*, 5644. Piecuch, P.; Kowalski, K.; Pimienta, I. S. O.; McGuire, M. J. *Int. Rev. Phys. Chem.* **2002**, *21*, 527.

(83) Miliordos, E.; Mavridis, A. *J. Chem. Phys.* **2008**, *128*, 144308.

(84) Bhaskaran-Nair, K.; Demel, O.; Pittnera, J. *J. Chem. Phys.* **2008**, *129*, 184105.

(85) Martin, J. M. L. *J. Chem. Phys.* **1992**, *97*, 5012. Martin, J. M. L.; Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843.

(86) Huber, K. P.; Herzberg, G. *Molecular Structure and Molecular Spectra. IV. Constants of Diatomic Molecules*; Van Nostrand-Reinhold: New York, 1979.

(87) Fernando, W. T. M. L.; Bernath, P. F. *J. Mol. Spectrosc.* **1991**, *145*, 392.

CT900452N

JCTC Journal of Chemical Theory and Computation

A First Principles Development of a General Anisotropic Potential for Polycyclic Aromatic Hydrocarbons

Tim S. Totton,[†] Alston J. Misquitta,*[,‡] and Markus Kraft[†]

Department of Chemical Engineering and Biotechnology, University of Cambridge, New Museums Site, Pembroke Street, Cambridge CB2 3RA, United Kingdom, and Department of Physics, Cavendish Laboratory, University of Cambridge, J J Thomson Avenue, Cambridge, CB3 0HE, United Kingdom

Received September 15, 2009

Abstract: Standard empirical atom−atom potentials are shown to be unable to describe the binding of polycyclic aromatic hydrocarbon (PAH) molecules in the variety of configurations seen in clusters. The main reason for this inadequacy is the lack of anisotropy in these potentials. We have constructed an anisotropic atom−atom intermolecular potential for the benzene molecule from first principles using a symmetry-adapted perturbation theory based on density functional theory (SAPT(DFT)), interaction energy calculations and the Williams−Stone−Misquitta method for obtaining molecular properties in distributed form. Using this potential as a starting point, we have constructed a transferable anisotropic potential to model intermolecular interactions between PAHs. This new potential has been shown to accurately model interaction energies for a variety of dimer configurations for four different PAH molecules, including certain configurations which are poorly modeled with current isotropic potentials. It is intended that this potential will form the basis for further work on the aggregation of PAHs.

1. Introduction

Polycyclic aromatic hydrocarbon (PAH) molecules have often been invoked as intermediates in the chemistry of soot formation and growth.[1] The presence of stacked PAH molecular structures in experimental high-resolution transmission electron microscopy (HRTEM) images of soot particles[2−4] has led some to suggest that the intermolecular binding of PAH molecules may be responsible for particle inception. This hypothesis has provoked a large number of theoretical studies on the stability and the relative orientation of PAH molecules present in dimers and larger stacks in flame environments.[1,5−9] Currently, many numerical simulations of soot formation in flames consider the dimerization of molecules as small as pyrene ($C_{16}H_{10}$)[9,10] to be the particle inception step, however, the validity of this assumption is still debated.[11]

The aggregation of PAH molecules has traditionally been modeled using atom−atom potentials, which approximate the total interaction energy, U, as sum over all pairwise atomic interactions between molecules:

$$U = \sum_{\mathrm{A}} \sum_{\mathrm{A<B}} \sum_{a\in \mathrm{A}} \sum_{b\in \mathrm{B}} U_{ab}(R_{ab}, \Omega_{ab}) \qquad (1)$$

Here $U_{ab}(R_{ab},\Omega_{ab})$ denotes an atom−atom interaction potential. The indices A and B are for molecules, and the indices a and b run over all the atomic sites within these molecules. In general, the interaction potential depends upon the atom−atom separation, R_{ab}, and the relative molecular orientation, described in some way by Ω_{ab}. Often, however, orientational dependence is removed as a simplification, and such potentials are 'isotropic', i.e., the atoms in a molecule are considered to be spherically symmetric. Previous studies into the intermolecular chemistry of PAHs have been largely based on computationally convenient model potentials, such as isotropic Lennard-Jones 12-6 (eq 2) and exp-6 potentials (eq 3).[1,12,13] Explicit electrostatic models are often added to these forms, the simplest being based upon partial atom-centered point charges, q_a and q_b (eq 4).

* Corresponding author. E-mail: am592@cam.ac.uk.

[†] Department of Chemical Engineering and Biotechnology, University of Cambridge.

[‡] Department of Physics, University of Cambridge.

10.1021/ct9004883
Published on Web 02/12/2010

$$U_{ab}^{\mathrm{LJ}} = 4\varepsilon_{ab}\left[\left(\frac{\sigma_{ab}}{R_{ab}}\right)^{12} - \left(\frac{\sigma_{ab}}{R_{ab}}\right)^{6}\right] \quad (2)$$

$$U_{ab}^{\exp-6} = B_{ab}\exp(-C_{ab}R_{ab}) - \frac{A_{ab}}{R_{ab}^{6}} \quad (3)$$

$$U_{ab}^{\mathrm{elst}} = \frac{q_a q_b}{R_{ab}} \quad (4)$$

In recent years, with the advance of computational power, the theoretical understanding of intermolecular interactions has developed significantly yet empirical potentials have remained largely unchanged. Current isotropic literature potentials have typically been parametrized to be applicable to a wide range of organic molecules.[14−17] Such potentials are typically parametrized with heats of sublimation and crystallographic data and, while transferable, are only accurate for the configurations they were parametrized for, and often fail at others. For example, consider two widely used potentials: the Williams W99 potential[15,16] based on the exp-6 form and the 12-6 Lennard-Jones potential,[14] both including point charges (the latter parametrized from an earlier form of the Williams potential). The performance of these isotropic potentials has been examined for naphthalene and anthracene dimer orientations shown in Figures 1 and 2. In Figures 3 and 4, we show cross sections of the potential energy surface at these orientations, where separation, R, is between the centers of mass of the monomers. The reference energies are taken from ab initio SAPT(DFT) calculations performed by Podeszwa and Szalewicz.[18]

The W99 potential performs remarkably well for stacked PAH geometries, while the Lennard-Jones plus point charges potential tends to overestimate well depths by 5−10 kJ mol^{-1}. However, both potentials show substantial errors for the T-shape dimer with an error in equilibrium separation of 0.3−0.4 Å and with an underbinding of as much as 7 kJ mol^{-1} in the case of the W99 potential.

Being isotropic, the potentials cannot accurately model the atom−atom interactions where there is significant anisotropy in the electron distribution around constituent atoms, such as in PAHs where there is significant π-bonding. These potentials also suffer from being required to possess too large a degree of transferability to make them sufficiently accurate for the specific system of interest, and development of anisotropic potentials empirically is precluded due to insufficient experimental data.

To accurately model dimers in all orientations either accurate ab initio methods must be used directly (on-the-fly methods) or new anisotropic atom−atom potentials are required, parametrized using ab initio results. Currently, most ab initio methods are prohibited due to high computational expense. Density functional theory (DFT) is the only method which is computationally feasible, but currently there are no practical and quantitative functionals which correctly predict intermolecular dispersion energies. In reality, the size of PAH systems (which can consist of many hundreds of atoms) and the complexity of the calculations restrict us to using model atom−atom potentials. In the context of PAHs and investigating soot structure at a molecular level, there are a number of requirements for a potential:

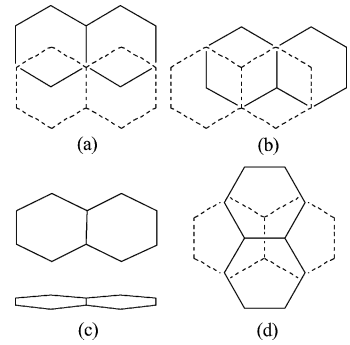


Figure 1. Naphthalene: (a) Slipped-parallel, symmetry C_{2h}; (b) Graphite-type, symmetry C_i; (c) T-shape, symmetry C_{2v}; (d) Crossed, symmetry D_{2d}.

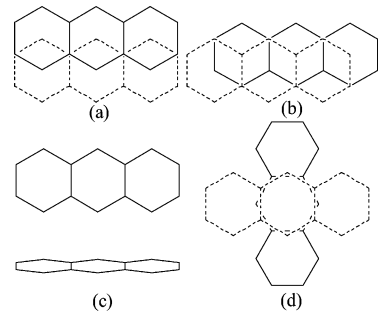


Figure 2. Anthracene: (a) Slipped-parallel, symmetry C_{2h}; (b) Graphite-type, symmetry C_i; (c) T-shape, symmetry C_{2v}; (d) Crossed, symmetry D_{2d}.

• Accuracy: The potential is expected to be accurate for the variety of dimer configurations which are expected to be sampled in a flame environment. In particular, the potential must correctly predict barriers on the potential energy surface (PES) of the molecular cluster.

• Transferability: In a flame environment, typically large ensembles of different PAH molecules exist of varying size (C_6−C_{400}),[20] and consequently, it is very important that any potential developed can be easily transferable to different PAHs.

• Simplicity: Large PAH clusters need to be studied requiring extensive calculations. This will limit the functional complexity of the potential expression used to model interactions.

In the context of PAHs, there have been several studies using ab initio methods. For example, coupled cluster calculations at CCSD(T) level have been used to study naphthalene dimers,[21] while MP2 level calculations have been used to obtain dimer interaction energies for various PAHs.[10] However, Møller−Plesset perturbation theory is inadequate to study intermolecular interactions between systems with a significant amount of π-bonding (such as PAH clusters). Compared to the more reliable CCSD(T) calculations, MP2 calculations have been shown to considerably overestimate attraction between molecules, in some cases by almost a factor of 2,[21−23] throwing into doubt some earlier studies of soot particle inception.[10] However, CCSD(T) is computationally demanding and is not ideally suited for potential development due to the inability to decompose the overall interaction energy into physically significant contributions. This makes it hard to parametrize

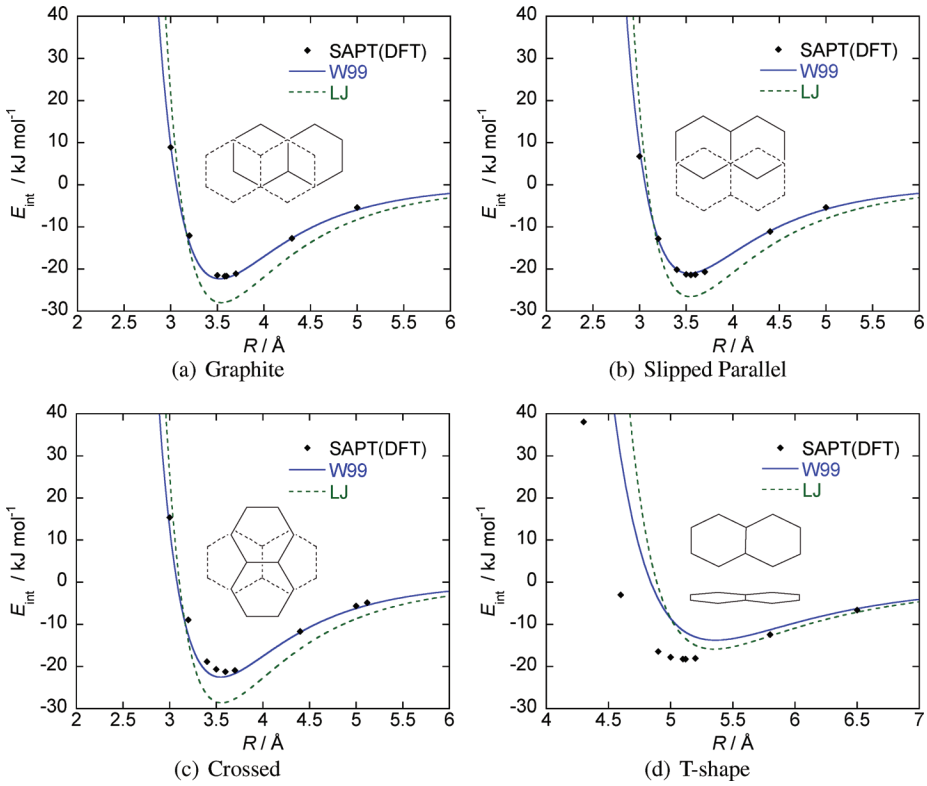


Figure 3. Comparison of isotropic model potentials with SAPT(DFT) energies for different naphthalene dimers. A key to the geometries is given in Figure 1. Model potential energies have been calculated using the Orient[19] program.

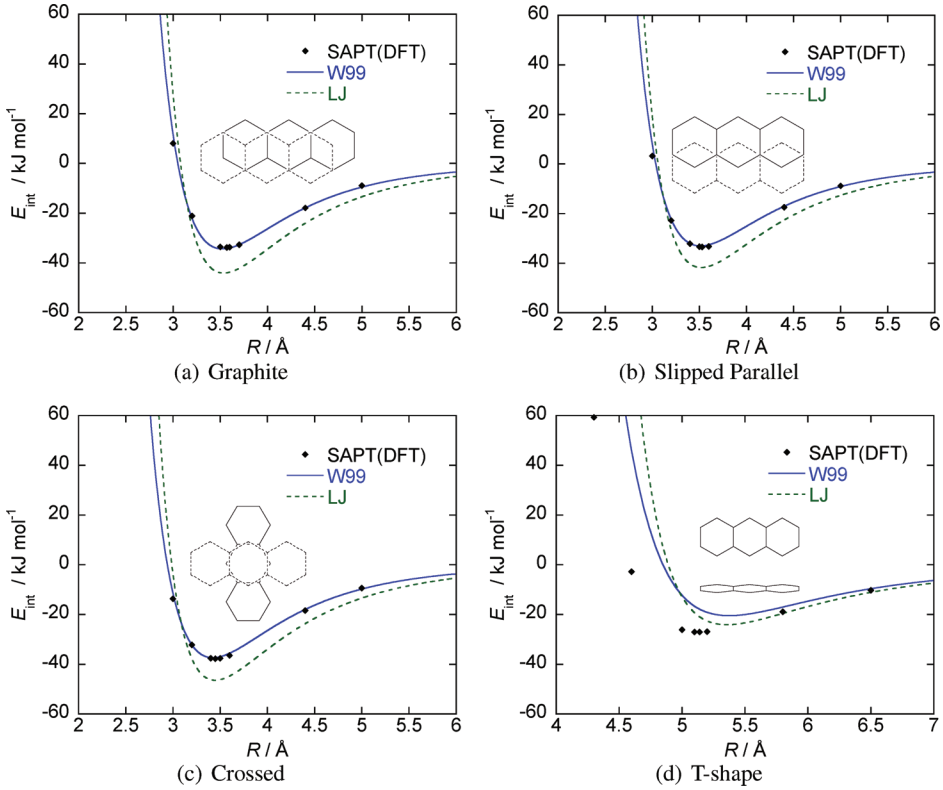


Figure 4. Comparison of isotropic model potentials with SAPT(DFT) energies for different anthracene dimers. A key to the geometries is given in Figure 2. Model potential energies have been calculated using the Orient[19] program.

analytic potentials comprised of multiple terms, where each term describes a different interaction, such as the dispersion or repulsion.

By contrast, intermolecular perturbation theory provides an ideal framework for the development of model potentials because it provides the interaction energy as a sum of

physically significant contributions. This allows the separate parametrization of different terms representing different interactions within multi-term model potentials. The development of symmetry-adapted perturbation theories has enabled both long- and short-range interactions to be accurately calculated, and the recent development of SAPT(DFT)[24−31] has made possible highly accurate studies of intermolecular interactions at a level comparable to CCSD(T),[32,33] with modest computational resources.

This methodology has already been used to develop intermolecular potentials. Misquitta et al.[34] have developed an anisotropic potential to predict the crystal structure of the 1,3-dibromo-2-chloro-5-fluorobenzene ($C_6BrClFH_2$) molecule, giving results in excellent agreement with experiment. Similarly, a potential derived from SAPT(DFT) interaction energies has been used to study the potential energy surface of the cyclotrimethylene trinitramine (RDX) dimer.[35] A benzene potential has been constructed using SAPT(DFT) energy calculations of 491 dimer geometries.[33] However, in addition to the usual atomic sites, this potential also contains off-atomic sites, and it is difficult to see how the parameters for off-atomic sites can be transferred easily to larger PAHs.

The potential form we have chosen to use is

$$U_{ab} = \underbrace{G\exp\left[-\alpha_{ab}\left(R_{ab}-\rho_{ab}(\Omega_{ab})\right)\right]}_{short-range} \underbrace{-f_6(R_{ab})\frac{C_{6,\mathrm{iso}}}{R_{ab}^6}+E_{\mathrm{elst}}(\mathrm{model})}_{long-range}, \qquad (5)$$

where the first term is a Born−Mayer term describing short-range interactions, the second is an isotropic, damped dispersion term, and the third term is an appropriate electrostatic model. These terms will be described in detail later. This form of the potential remedies two of the major deficiencies of the exp-6 potential in eq 3: (i) the short-range term now includes a shape-function, ρ_{ab}, that models the anisotropy of the interacting sites through a dependence on the relative orientation, Ω_{ab}, of the two sites, and (ii) the singularity in the dispersion term is removed by the damping function, $f_6(R_{ab})$ (defined later). A more elaborate potential functional form could have been chosen, for example, we could have included higher order terms and anisotropy in the dispersion model and the anisotropy in the damping function, but the resulting potential would be unnecessarily computationally demanding and would not be usable in most simulation programs without a significant degree of modification.

We begin this article with a description of the methods we have used to parametrize this potential for benzene. We fit energies and molecular properties to the best ab initio data provided by SAPT(DFT) and use the Williams−Stone−Misquitta (WSM) method[36−40] for determining distributed frequency-dependent polarizabilities needed for the dispersion model (Section 2). In order to keep the parameters physical, we have used a multistage fitting procedure to obtain the parameters for the short-range part of the potential (Section 3). The resulting parameter set then acts as a starting point for the generalization of the potential to larger PAHs (Section 4). In this stage, we have used SAPT(DFT) interaction energies calculated by Podeszwa and Szalewicz[18] for dimers of naphthalene ($C_{10}H_8$), anthracene ($C_{14}H_{10}$), and pyrene, ($C_{16}H_{10}$) in a variety of configurations. Finally, we conclude with an assessment of the accuracy of SAPT(DFT) energies and possible directions for future applications of the potential (Section 5).

2. Constructing the Benzene Dimer Potential

The basic strategy for constructing an analytic potential for molecules consisting of more than two atoms has been described in a recent review.[41] The potential is logically separated into long- and short-range parts (eq 5). The long-range part depends upon molecular properties, such as multipole moments, polarizabilities, and dispersion coefficients. Long-range polarization, or induction, is expected to be weak in molecules which do not possess strong multipole moments, such as PAHs. We have, therefore, omitted an explicit induction term in our model potentials. The short-range energies include the exchange-repulsion, the penetration energies (see below), and the second-order induction effects (which make a small but significant contribution). These energies all decay exponentially with increasing separation. We have fitted the parameters of the exponential terms via the density overlap model, using the procedure described in ref 34 and outlined below.

2.1. Molecular Geometry and Basis Sets. The geometry of benzene was obtained by in vacuo optimization using DFT with the B3LYP functional and the 6-31G* basis set with the Gaussian03[42] program. The molecule was assumed to be rigid. Calculated atomic coordinates are provided in the Supporting Information. The D_{6h} symmetry of benzene allows us to identify just two unique atom types: a carbon and a hydrogen. Symmetry was imposed during the calculation of the distributed properties and the subsequent fitting process for the Born−Mayer parameters.

Interaction energies and molecular properties have been calculated using the CamCASP[43] program from molecular wave functions obtained using the Dalton[44] program. The molecular wave functions were calculated with the asymptotically corrected PBE0[45] exchange−correlation functional and the Sadlej-pVTZ[46] basis. We have used the Tozer−Handy asymptotic correction[47,48] with a vertical ionization potential of 0.3397 a.u., obtained from ref 49. The linear-response DFT calculations needed for second-order SAPT(DFT) energies were performed using a hybrid adiabatic local density approximation (LDA) and coupled Hartree−Fock kernel.[27,38]

All calculations have used the Sadlej-pVTZ basis[46] set in two basis types: (i) a 'monomer-centered' (MC) basis, which includes basis functions on atomic sites only, and (ii) a 'monomer-centered-plus' (MC+) basis type, which additionally includes basis functions placed in the bonding region between the two molecules and on the atomic sites of the partner molecule. The MC type of basis was used for molecular properties and for the first-order interaction energies (the electrostatic and exchange-repulsion energies) and density-overlaps. This type of basis is not suitable for calculations of the second-order interaction energies which are slow to converge with basis set.[50] For these, we have used the MC+ type of basis with a 3s2p1d basis set for the mid-bond functions, placed at a position determined by a generalization of the weighting scheme described in ref 51.

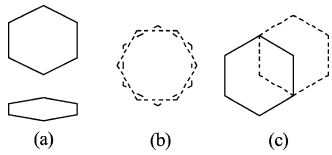


Figure 5. Benzene: (a) T-shape, symmetry C_{2v}; (b) Crossed, symmetry D_{6d}; (c) Slipped-parallel, symmetry C_{2v}.

The CamCASP program uses density fitting in the evaluation of interaction energies, molecular properties, and density overlaps. We used two kinds of auxiliary basis sets in our calculations: (i) the aug-cc-pVTZ auxiliary basis[52] has been used for the calculations of molecular properties and SAPT(DFT) energies with the MC+ basis type, and (ii) the smaller JK-TZVPP auxiliary basis[52] has been used for the calculation of the density overlap and the first-order SAPT(DFT) energies used in the first part of the fitting process.

2.2. SAPT(DFT) Dimer Energies. SAPT(DFT) interaction energies for the benzene dimer were calculated at a variety of dimer configurations so as to model the exchange−repulsion, penetration, and induction energies as well as to provide a set of reference dispersion energies for the assessment of our dispersion models. Based upon an earlier study,[27,34] we used the following formulation of the SAPT(DFT) interaction energy:

$$E_{\mathrm{int}}^{(2)} = E_{\mathrm{elst}}^{(1)}(\mathrm{KS}) + E_{\mathrm{exch}}^{(1)}(\mathrm{KS}) + E_{\mathrm{ind,tot}}^{(2)} + E_{\mathrm{disp,tot}}^{(2)} \quad (6)$$

where, $E_{\mathrm{elst}}^{(1)}(\mathrm{KS})$ and $E_{\mathrm{exch}}^{(1)}(\mathrm{KS})$ are the first-order electrostatic and exchange−repulsion energies, $E_{\mathrm{ind,tot}}^{(2)}$ is the total induction energy defined as sum of the polarization and exchange contributions,[27] $E_{\mathrm{ind,pol}}^{(2)} + E_{\mathrm{ind,exch}}^{(2)}$, and likewise, $E_{\mathrm{disp,tot}}^{(2)}$ is the total dispersion energy defined as $E_{\mathrm{disp,pol}}^{(2)} + E_{\mathrm{disp,exch}}^{(2)}$. The second-order terms are calculated using Kohn−Sham linear response theory. Terms of third- and higher-order in the interaction operator have been omitted, as these are not expected to be significant for systems without hydrogen bonds.[33,38]

It is important to select the dimers so as to get a uniform coverage of the space of physically important configurations. We have done this by keeping one of the molecules of the dimer fixed and centered at the origin and by translating and rotating the other using the following algorithm:[34]

• Using a Sobol quasi-random sequence, generate a random direction vector for the translation and, using Shoemake's uniform distribution algorithm,[53] generate a quaternion for rotation.

• Starting with both molecules centered at the origin, rotate one using the quaternion. Using the standard van der Waals radii,[54] determine the distance of van der Waals contact R_0 along the direction vector.

• Translate the rotated molecule along the direction vector by a few (1−5) randomly selected distances chosen to lie between $R_0 - \Delta R_{\mathrm{min}}$ and $R_0 + \Delta R_{\mathrm{max}}$. We have used ΔR_{min} and ΔR_{max} to be 1.5 and 1.2 a.u., respectively.

In addition to 500 benzene dimer configurations selected, using the algorithm described above, we have used an additional 27 energies that were calculated at specific dimer orientations shown in Figure 5. Here the slipped-parallel and crossed configurations represent stacked dimer configurations, and the interaction energy has been calculated at a set of interplanar spacings. Likewise, the T-shaped configuration energies have been calculated at a set of separations of the monomer centers of mass.

2.3. Molecular Properties. Ideally, when the effect of electron density overlap is negligible, we would describe the electrostatic interaction energy using a distributed multipole model.[55,56] Such models have had a lot of success in organic crystal structure prediction.[57] However, for the benzene dimer, we have found that a simple point charge model suffices, and the higher ranking multipole moments are not so important. This is probably because of the absence of strong directional moments, such as those seen in hydrogen-bonding complexes. Additionally, there are few simulation programs that can use distributed multipole models. Consequently, we have used a distributed point charge model to describe the electrostatic interaction at long range. This model was calculated with the Gaussian03[42] program using the PBE0 functional and Sadlej-pVTZ basis using the Merz−Singh−Kollman scheme,[58] which fits the molecular electrostatic potential to a set of atom-centered point charges. In principal, the electrostatic term, which varies as $1/R_{ab}$ should be damped at short range to avoid the divergence as $R_{ab} \rightarrow 0$. In practice however, the low power of R_{ab} means that this divergence is manifest only at very small separations, and hence, damping can be ignored in many cases.

Frequency-dependent polarizabilities are needed to calculate the dispersion coefficients with which we model the second-order dispersion energies at intermolecular separations where orbital overlap effects can be neglected. These polarizabilities need to be distributed in order to ensure rapid convergence with rank of the multipole expansion; in fact, even for a molecule the size of benzene, the single center multipole expansion will not converge for the physically important dimer configurations. Distributed frequency-dependent polarizabilities of ranks one, two, and three for carbon and hydrogen atoms have been obtained using the WSM method. This distribution method has been shown to result in models which exhibit very good convergence properties, while resulting in a physically meaningful partitioning of the molecular properties.

2.4. Dispersion Models. In the multipole expansion, the second-order dispersion energy between two molecules, A and B, is given by[59]

$$E_{\mathrm{disp,d}}^{(2)} = -\sum_{a\in \mathrm{A}}\sum_{b\in \mathrm{B}}\left(f_6(R_{ab})\frac{C_6^{ab}}{R_{ab}^6} + f_7(R_{ab})\frac{C_7^{ab}}{R_{ab}^7} + f_8(R_{ab})\frac{C_8^{ab}}{R_{ab}^8} + \ldots\right) \quad (7)$$

where the C_n^{ab} are the dispersion coefficients which are orientationally dependent for non-spherically symmetric sites. For interactions between spherically symmetric sites, the C_n^{ab} coefficients contain no angular dependence and terms that are odd in n are zero. The $f_n(R_{ab})$ are damping functions,[59] which are needed to remove the divergence of the expansion at small R_{ab}. Here, we use the Tang−Toennies damping functions:[60]

$$f_n(R_{ab}) = 1 - \exp(-\beta R_{ab}) \sum_{k=0}^{n} \frac{(\beta R_{ab})^k}{k!} \quad (8)$$

with an isotropic-damping parameter of $\beta = 2(2I)^{1/2}$, where I is the vertical ionization energy in a.u.[40] For benzene, using the value of I presented above, we obtain $\beta = 1.6485$ a.u.

The WSM method allows the calculation of a variety of dispersion models from the simple isotropic C_6 model to the very elaborate anisotropic C_{12} model[40] (a C_n model includes all terms from C_6 to C_n). We have calculated isotropic and anisotropic C_6, C_{10}, and C_{12} models using the CamCASP[43] program. In Figure 6, we present dispersion energies from these models calculated with the Orient program and displayed against SAPT(DFT) total dispersion energies, $E^{(2)}_{\mathrm{disp,tot}}$.

2.4.1. Refining the Isotropic C_6 Dispersion Model. In practice, elaborate anisotropic dispersion models are computationally demanding, and few simulation programs are able to use them. We have, therefore, created an *effective* damped isotropic C_6 model in the manner described in ref 40.

The scatter plot shown in Figure 6 shows that deviation of the model dispersion energies from the SAPT(DFT) energies is approximately linear for all the dispersion models. It, thus, becomes possible to introduce a scaling factor by which the $C_{6,\,\mathrm{iso}}$ model dispersion energies can be scaled to recover more accurately the SAPT(DFT) energies. In order to find the scaling coefficients, a function of the following form was minimized

$$\Lambda = \sum_i w_i \left[E^i_{\mathrm{disp,tot}} + \xi \sum_{a\in \mathrm{A}, b\in \mathrm{B}} \frac{f_6(R_{ab}) C^{ab}_{6,\mathrm{iso}}}{R^6_{ab}} \right]^2 \quad (9)$$

where i labels the configurations, and $E_{\mathrm{disp,tot}}$ is the total SAPT(DFT) dispersion energy. The coefficient, ξ, is determined by a least-squares fit, and w_i is a weight, which will generally be energy dependent. In the general case, the scaling coefficient would depend on the atom pairs, but in this work, the simplest possible fit has been considered: all configurations are weighted equally, and a single constant of proportionality is used. The appropriate scaling factor for the damped isotropic C_6 model was found to be 1.372 for the physically significant dispersion energies defined by the range from -20 to 0 kJ mol^{-1}. The root-mean-square (rms) error for the damped and scaled $C_{6,\,\mathrm{iso}}$ model over this range is only 0.47 kJ mol^{-1}. At small separations when total dispersion energies are lower than -20 kJ mol^{-1}, the dispersion energy is overestimated, but at such short ranges, repulsive interactions are expected to dominate. The dispersion is also overestimated at large separations, although it should be noted that this is true of any *effective* C_6 representation of dispersion using only atomic sites.[40]

3. Short-Range Energies

The short-range energy is defined as the difference in the total interaction energy (eq 6) and the energies calculated with the multipole expansions for the electrostatics and the dispersion. As mentioned above, since the long-range induction is expected to be weak in systems of PAH molecules, we have not included a multipole expansion for the induction energy. Therefore, we define the short-range energy is as:

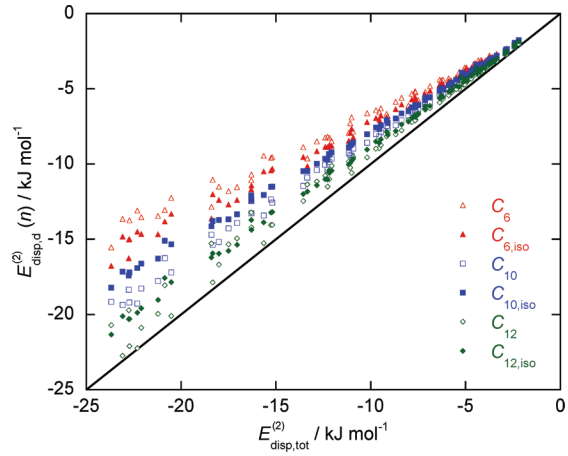


Figure 6. Dispersion energies for the benzene dimer. Scatter plot of dispersion energies calculated using the damped dispersion models represented by $E^{(2)}_{\mathrm{disp,d}}(n)$ against $E^{(2)}_{\mathrm{disp,tot}}$ calculated using SAPT(DFT). The dispersion models presented are anisotropic unless given the suffix 'iso', in which case they are isotropic.

$$E^{(2)}_{\mathrm{sr}} = (E^{(1)}_{\mathrm{elst}} - E^{(1)}_{\mathrm{elst}}(\mathrm{ESP})) + E^{(1)}_{\mathrm{exch}} + E^{(2)}_{\mathrm{ind,tot}} = E^{(1)}_{\mathrm{elst,pen}} + E^{(1)}_{\mathrm{exch}} + E^{(2)}_{\mathrm{ind,tot}} \quad (10)$$

where $E^{(1)}_{\mathrm{elst}}(\mathrm{ESP})$ is the electrostatic energy calculated using the point charge model, $E^{(2)}_{\mathrm{disp,d}}(C_{6,\,\mathrm{iso}})$ is the dispersion energy calculated using the *effective* damped $C_{6,\,\mathrm{iso}}$ model, and $E^{(1)}_{\mathrm{elst,pen}}$ is the electrostatic penetration energy defined implicitly above. The induction is included as a short-range energy because, in the absence of strong permanent multipoles, $E_{\mathrm{ind,tot}}$ is almost all due to orbital overlap effects. We could define a penetration-like contribution from the dispersion energy, $E^{(2)}_{\mathrm{disp,tot}} - E^{(2)}_{\mathrm{disp,d}}(C_{6,\,\mathrm{iso}})$, but this term is small and does not exhibit an exponential dependence with distance; consequently, we have omitted it here but have included it in the final stage of the fitting process.

Unlike the exchange−repulsion energy, $E^{(2)}_{\mathrm{sr}}$ is not always positive, as there will be configurations for which the penetration energies—which are generally negative—and the (negative) induction energies will be larger in magnitude than the corresponding exchange−repulsion energies. Nevertheless, this happens at very few dimer configurations, and when it does, $E^{(2)}_{\mathrm{sr}}$ tends to be small in magnitude. We have, therefore, omitted such configurations. Furthermore, since the penetration energy and exchange−repulsion energies both arise from the overlap of the molecular wave functions, they both exhibit an exponential dependence on intermolecular separation. If we assume the same distance dependence for all terms in eq 10, we can then fit the positive values of $E^{(2)}_{\mathrm{sr}}$ to the Born−Mayer term of eq 5:

$$\sum_{a\in \mathrm{A}} \sum_{b\in \mathrm{B}} G\exp[-\alpha_{ab}(R_{ab} - \rho_{ab}(\Omega_{ab}))] \quad (11)$$

The hardness of the interaction is described by α_{ab}, and G is a constant energy unit taken to be 10^{-3} a.u. The shape

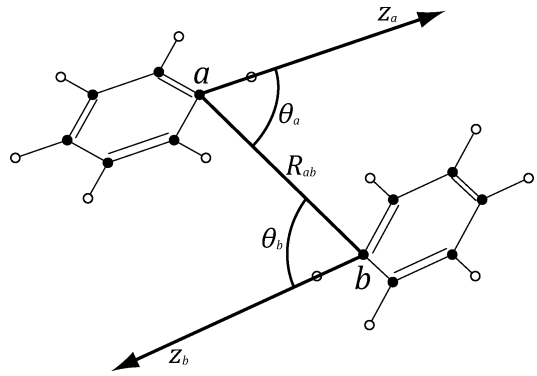


Figure 7. Schematic showing the axis system used to define short-range anisotropy. The local *z*-axes for both the carbon atoms and the attached hydrogen atoms point radially outward along the carbon−hydrogen bonds.

functions, $\rho_{ab}(\Omega_{ab})$, have been assumed to be the sum of the shape functions of the individual sites:[59]

$$\rho_{ab}(\Omega_{ab}) = \rho^a(\theta_a) + \rho^b(\theta_b) \tag{12}$$

where

$$\rho^a(\theta_a) = \rho^a_{00} + \rho^a_{10}\cos\theta_a + \frac{1}{2}\rho^a_{20}(3\cos^2\theta_a - 1) \tag{13}$$

and similarly for $\rho^b(\theta_b)$. Here the angle θ_a defines the angle between the site−site vector from a to b and the z-axis in the local axis system of site a. We have used an approximate axial symmetry at each atomic site with the z-axis in the local axis system pointing radially outward from the carbon to the attached hydrogen (Figure 7). In this work, the shape function for carbon included terms to rank two, and for hydrogen, where anisotropy is less important, only terms up to rank one were found to be necessary.

Rather than attempt a direct fit of eq 11 to the short-range energies, we have performed the fitting in several stages. One of the reasons for this is numerical stability. In general, due to the highly coupled nature of the parameters in eq 11, a direct fit tends to result in an unphysical parameter set and is, therefore, best avoided, particularly if the parameters are to be *transferred* to other, similar, systems. Another reason is one of computational efficiency; if a first approximation to the parameters can be obtained using low-level ab inito data obtained at a low computational cost and subsequently improved using a smaller set of higher quality ab inito data, then the overall computational cost of the fitting process is reduced.

3.1. Stage 1: Fit to $E^{(1)}_{sr} = E^{(1)}_{exch} + E^{(1)}_{elst,pen}$. In the first stage, we have fitted the Born−Mayer terms to the first-order contribution to the short-range energy:

$$E^{(1)}_{sr} = E^{(1)}_{exch} + E^{(1)}_{elst,pen} \tag{14}$$

Since the first-order energies constitute the major part of the total short-range energy, the resulting Born−Mayer parameters are a good starting point for the next stage of the fitting process. Additionally, because these terms are computationally inexpensive, we can calculate $E^{(1)}_{sr}$ at a fairly large number of dimer configurations with a modest amount of computational resource. We have computed $E^{(1)}_{sr}$ at 500 benzene configurations using the CamCASP[43] program. These calculations were performed using the MC Sadlej-pVTZ basis together with the JK-TZVPP auxiliary basis.

We have used the method described in ref 34 to fit eq 11 to $E^{(1)}_{sr}$. As pointed out above, the parameters in eq 11, which is a sum of exponential terms, tend to be highly coupled, and a direct fit often leads to an unphysical parameter set. It would be more appropriate to fit the exponential terms in eq 11 *individually*, for each pair of sites, but to achieve this, we need to partition the short-range energy $E^{(1)}_{sr}$ into contributions from pairs of atoms. While this partitioning cannot be rigorously defined, a reasonable break-up of the short-range energy can be obtained through the density overlap model[61] in a manner outlined below and described in more detail in refs 34 and 41.

The density overlap model postulates that the short-range energy is nearly proportional to the overlap between the molecular electron densities. The short-range energy is generally taken to be the exchange−repulsion energy, but here, we additionally include the first-order electrostatic penetration energy. Therefore, we have

$$E^{(1)}_{sr} = K_0 S^{\gamma}_{\rho} \tag{15}$$

where, K_0 and γ are constants, and if ρ^X_e is the electron density of molecule X, then the density overlap is defined as $S_\rho = \int \rho^A_e(\mathbf{r})\rho^B_e(\mathbf{r})d^3\mathbf{r}$. For the asymptotically correct densities we have used here, the exponent γ has been shown[62] to be exactly 1, so K_0 is the only free parameter in this model which can be determined by least-squares minimization.

Now, if we partition the electron density into atomic contributions, that is $\rho^A_e(\mathbf{r}) = \sum_{a\in A}\rho^a_e(\mathbf{r})$, then we can define a *distributed* density overlap $S^{ab}_\rho = \int \rho^a_e(\mathbf{r})\rho^b_e(\mathbf{r})d^3\mathbf{r}$, and, hence, a generalized form of the overlap model:[41,63,64]

$$E^{(1)}_{sr} = \sum_{a\in A, b\in B} E^{(1)}_{sr}(ab) = \sum_{a\in A, b\in B} K^{ab} S^{ab}_{\rho} \tag{16}$$

where K^{ab} are constants to be fitted, and the partitioned short-range energy $E^{(1)}_{sr}(ab)$ is implicitly defined through the above equation. The density partitioning is not unique and can be achieved in a variety of ways. We have used density fitting to achieve the partitioning, which is analogous to the Gaussian multipole method of Wheatley.[65] Details of weighting schemes and constraints used in the fitting process are described in ref 34.

Having obtained the atom−atom partitioned short-range energy $E^{(1)}_{sr}(ab)$, it is now relatively easy to fit the contributions of individual pairs of sites to a single Born−Mayer term from eq 11. These fits are well-defined and result in physically sensible values for the parameters.

The various stages of the above process were performed using the CamCASP[43] and Orient[19] programs, and the overall weighted rms error for the fitted energies was 0.82 kJ mol^{-1}. Symmetry was taken into account at all stages in the fitting process.

Table 1. Parameters of BAP in a.u.[a]

atom pair	$l_a\kappa_a$	$l_b\kappa_b$	ρ	α	C_6
CC	00	00	4.1780	1.8683	30.452
	10	00	0.2535		
	20	00	−2.0390		
CH	00	00	5.4242	1.7370	12.490
	00	10	−0.4663		
	10	00	0.1472		
	20	00	−0.1422		
HH	00	00	3.4400	1.5263	5.092
	10	00	0.3611		

[a] The pre-exponential factor, G, is 0.001 a.u., and the damping factor, β, is 1.6485 a.u. The C_6 coefficients quoted here include the scaling factor of 1.372 (see discussion in Section 2.4.1). The point charges used for carbon and hydrogen atoms are −0.1111 and 0.1111 a.u., respectively.

3.2. Stage 2: Fit to $E_{sr}^{(2)}$. We now refine the parameter set obtained in Stage 1, Section (3.1), against a smaller but higher-quality data set of short-range energies, this time using the second-order terms, as defined in eq 10. $E_{sr}^{(2)}$ was obtained using SAPT(DFT) energies calculated at the first 100 configurations used in Stage 1 and the further 27 specific dimers at the orientations shown in Figure 5. This time we used the much larger MC+ basis type.

The relaxation of the parameter set was performed using penalty functions of the form $(p_i - p_i^0)^2$, where p_i^0 is the anchor value obtained from Stage 1. In this way, the fit was refined while preventing the parameters from taking on non-physical values. The final choice of parameters to be relaxed and the weights given to the harmonic constraints were chosen with an element of empiricism, although values of α_{ab} and $\rho_{00,00}^{ab}$ were constrained more tightly than other parameters. The final fit had a weighted rms residual error of 0.75 kJ mol^{-1}.

3.3. Stage 3: Final Fit to Relax all Parameters. In the final step, we simultaneously refined the Born−Mayer parameters and the dispersion coefficients (with harmonic constraints imposed) to fit SAPT(DFT) energies calculated in the MC+ basis at the 127 geometries used in Stage 2, Section (3.2). The parameters in the electrostatic model were kept fixed during this step. The SAPT(DFT) energies were weighted so as to favor more negative energies to ensure the potential well was accurately fitted. The weighting scheme used has been adapted from that used in ref 66 and is given by

$$w_i = \Theta(E_{int}^i - E_0)\left(\frac{E_0}{E_{int}^i}\right)^2 + [1 - \Theta(E_{int}^i - E_0)]\exp[\eta(E_0 - E_{int}^i)] \quad (17)$$

where $\Theta(x)$ is the Heaviside step function. E_{int}^i are the SAPT(DFT) energies, and the parameters E_0 and η were set to 3 and 0.1 mol kJ^{-1}, respectively.

The parameters for the resulting benzene anisotropic potential (BAP) are given in Table 1. It should be noted that for the shape function coefficients, $\rho_{l_a\kappa_a, l_b\kappa_b}^{ab}$, of a given atom pair (i.e., C−C, C−H, or H−H) $\rho_{00,00}^{ab}$ is defined to be the sum of the rank 00 terms in eq 13, that is, $\rho_{00,00}^{ab} = \rho_{00}^a + \rho_{00}^b$, whereas the mixed-rank terms are defined as $\rho_{i0,00}^{ab} = \rho_{i0}^a$, and $\rho_{00,i0}^{ab} = \rho_{i0}^b$ for $i \in \{1, 2\}$.[59]

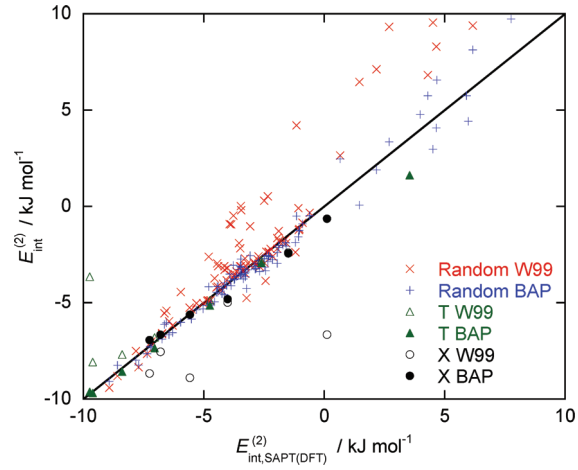


Figure 8. Comparison of the benzene anisotropic potential (BAP) with SAPT(DFT) energies calculated for 100 random benzene dimer orientations.

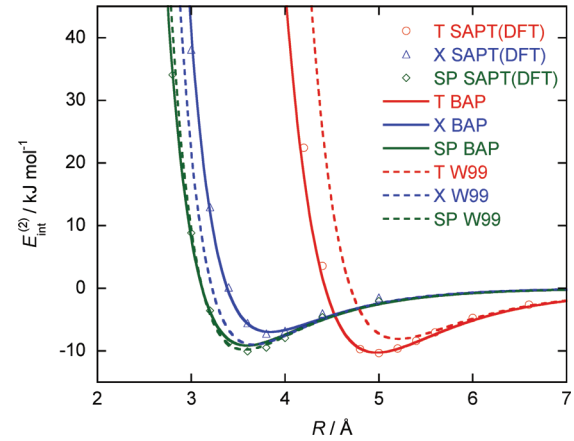


Figure 9. Comparison of BAP with SAPT(DFT) energies and W99 potential for T-shaped (T), crossed (X) and slipped-parallel (SP) dimer configurations.

The weighted rms residual energy $(E_{int}^i - E_{fit}^i)$ for the benzene anisotropic potential, when compared with the 127 SAPT(DFT) energies, was found to be 0.49 kJ mol^{-1}. Figure 8 shows the scatter of energies of the new potential compared to SAPT(DFT) energies for the random benzene dimer configurations and for some of the specific configurations chosen. For comparison, the W99 potential has been included, and the plot shows energies calculated with the new potential are noticeably less scattered. The scatter which remains for the new potential results is likely to be due to the damped C_6 isotropic dispersion model, which cannot be further improved without going to a more detailed model. Figure 9 shows a comparison of the new benzene potential with SAPT(DFT) energies and the W99 potential for the orientations shown in Figure 5. The new potential matches the SAPT(DFT) results in all configurations, especially the T-shaped configuration, where the W99 potential is notably poor.

The shape function shown in eq 12 imposes certain constraints on the parameters of the potential, e.g., $\rho_{10,00}^{HH} = \rho_{00,10}^{CH}$. However, these conditions were not imposed during the construction of the benzene anisotropic potential, and instead, shape functions for individual atoms were allowed

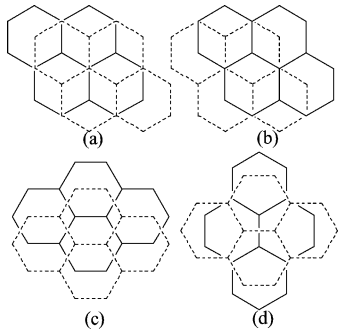


Figure 10. Pyrene: (a) Slipped-parallel L, symmetry C_{2h}; (b) Graphite-type, symmetry C_i; (c) Slipped-parallel S, symmetry C_{2h}; (d) Crossed, symmetry D_{2d}.

to vary depending on the specific atom pair considered. If the constraints are imposed, then the reduced flexibility of the functional form results in a poorer fit with a weighted rms residual of 0.96 kJ mol^{-1}.

4. Generalizing to Larger PAH Molecules

The shape function constraints given by eq 12 are probably inconsequential if the potential is restricted to a single system, but since they impose the idea of transferability,[59] they are needed if we wish to use the potential parameters on other, related, systems as well. So when generalizing the benzene potential to larger PAH molecules, we have imposed these constraints. This results in a potential with fewer parameters, but as shall be shown, it appears to perform remarkably well for the larger PAH dimers.

As can be seen from Figures 11−14, the BAP parameters transferred to the larger PAH molecules describe the interactions reasonably well. However, they are not entirely appropriate for a generalized transferable potential, as they do not satisfy the shape-function constraints. Consequently, we have used the benzene anisotropic potential parameters with these constraints imposed as the starting parametrization for the transferable anisotropic PAH potential. We have tuned these potential parameters against 111 SAPT(DFT) dimer energies, calculated by Podeszwa and Szalewicz,[18] for the naphthalene, anthracene and pyrene dimers at the orientations shown in Figures 1, 2, and 10. Molecular geometries were taken from ref 18, and ESP point charge models were calculated for each molecule with the Gaussian03[42] program in the same way as described for benzene. The geometries and partial atomic charges together with figures explaining the axes systems used in the larger PAH molecules are given in the Supporting Information. Also included in the Supporting Information are the files used to define the local axes systems in the Orient program.

The Orient program cannot simultaneously fit parameters to multiple types of molecular dimers, so an iterative scheme has been adopted. In this scheme, the initial benzene parametrization is used as the starting point for fitting the parameters for the naphthalene dimer; having obtained the new set of parameters, these now become the starting point for fitting to the anthracene dimer energies. This process is then continued, cycling through each set of dimer energies for each of the four PAH molecules. In order to converge to a parameter set, the harmonic constraints used in the fitting procedure were tightened after each iteration. Eventually the parameters are so tightly constrained that they could not be varied; this gave us the final parameter set. While this

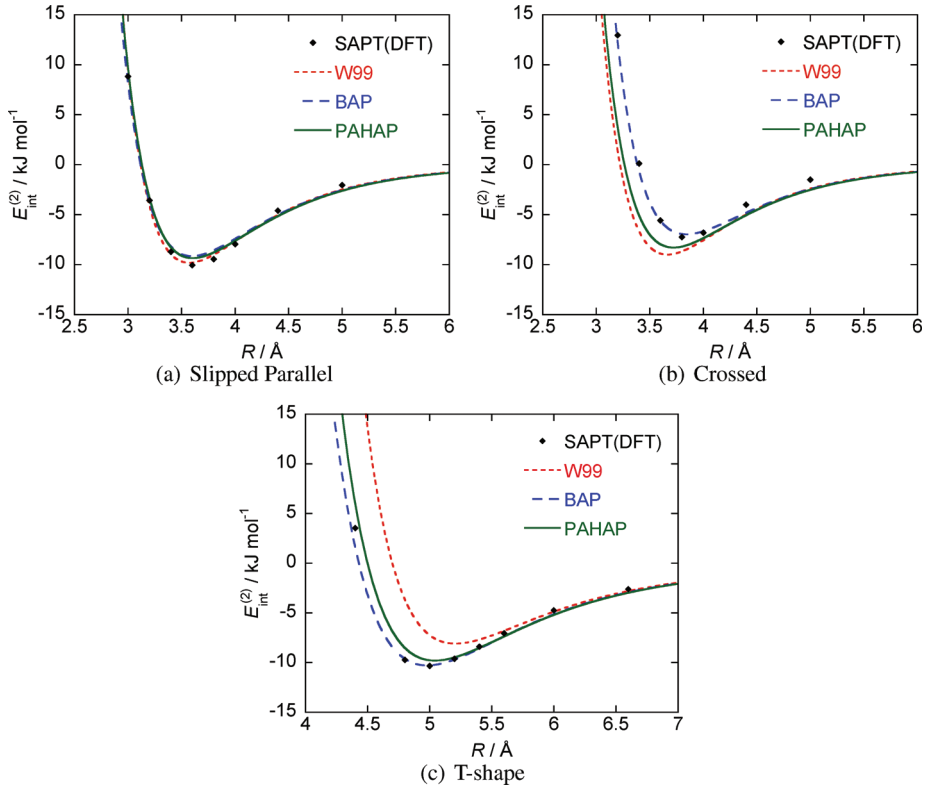


Figure 11. Comparison of the W99 potential, the BAP, and the PAH anisotropic potential (PAHAP) with SAPT(DFT) energies for benzene dimers. Model potential energies have been calculated using the Orient[19] program.

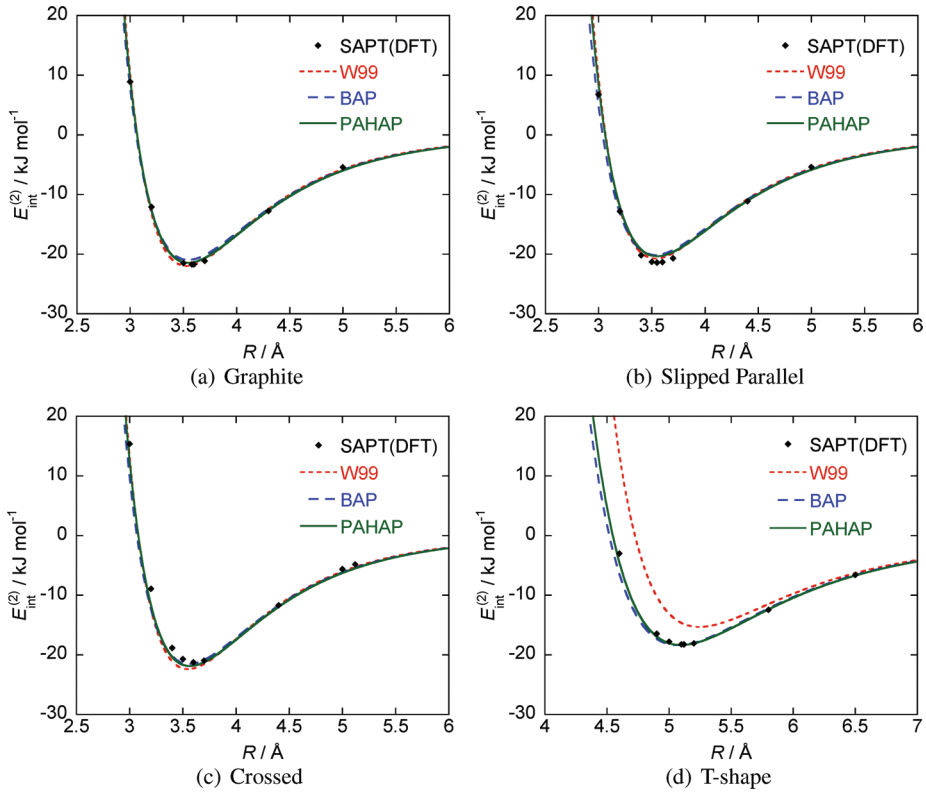


Figure 12. Comparison of the W99 potential, the BAP, and the PAHAP with SAPT(DFT) energies for naphthalene dimers. Model potential energies have been calculated using the Orient[19] program.

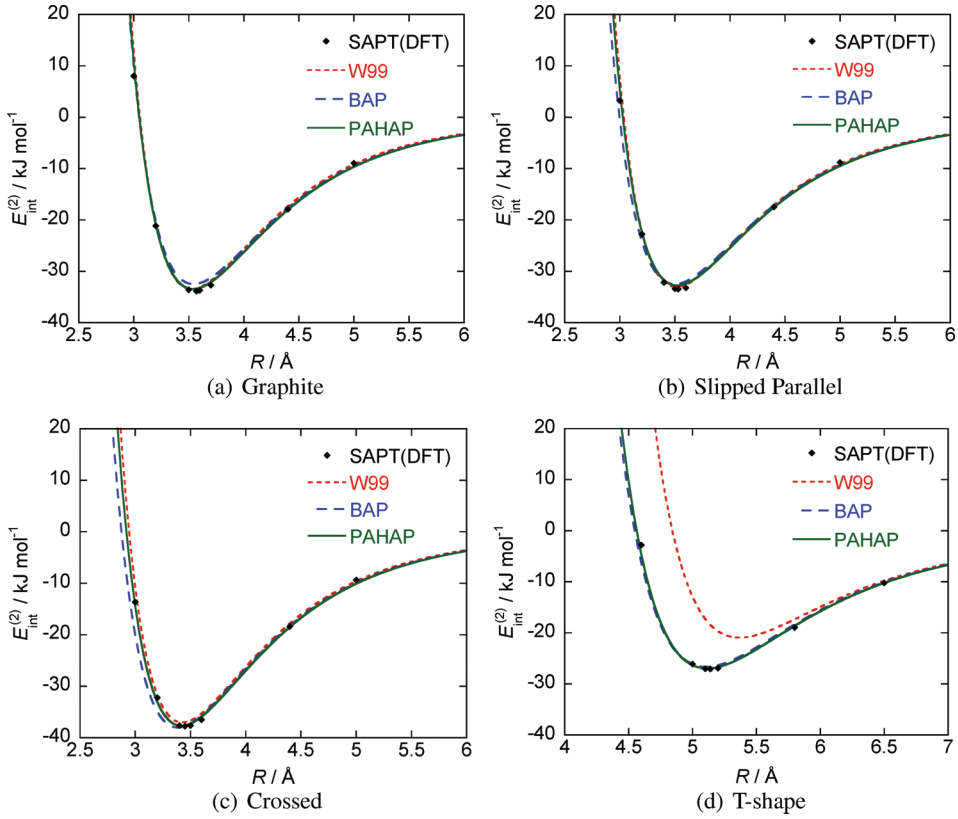


Figure 13. Comparison of the W99 potential, the BAP, and the PAHAP with SAPT(DFT) energies for anthracene dimers. Model potential energies have been calculated using the Orient[19] program.

procedure is by no means optimum, it has proved adequate and resulted in a generalized parameter set that is not only able to model the interactions of the larger PAH molecules but also the 127 benzene dimer geometries.

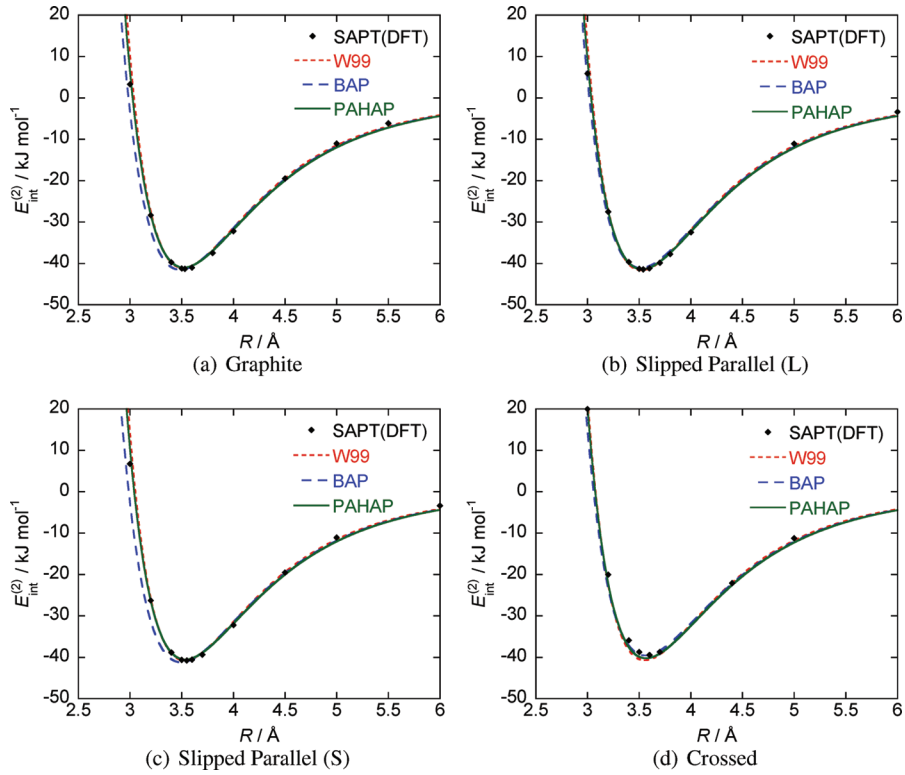


Figure 14. Comparison of the W99 potential, the BAP, and the PAHAP with SAPT(DFT) energies for pyrene dimers. Model potential energies have been calculated using the Orient[19] program.

Table 2. Parameters of PAH Anisotropic Atom−atom Potential in a.u.[a]

atom pair	$l_a\kappa_a$	$l_b\kappa_b$	ρ	α	C_6
C C	00	00	5.8147	1.8615	30.469
	10	00	0.0217		
	20	00	−0.2208		
C H	00	00	5.1505	1.7756	12.840
	00	10	−0.2718		
	10	00	0.0217		
	20	00	−0.2208		
H H	00	00	4.4862	1.4312	5.359
	10	00	−0.2718		

[a] The pre-exponential factor, G, is 0.001, and the damping factor, β, is 1.6485. The C_6 coefficients quoted here are effective coefficients and include an implicit scaling factor.

Table 2 shows the set of parameters obtained for the general PAH anisotropic potential (PAHAP) which, unlike the initial benzene parametrization, satisfy the constraints imposed by the shape function given in eq 12. A brief comparison of the new PAHAP shape functions with the equivalent W99 shape functions is given in the Supporting Information. From Figures 11−14, we see that the PAHAP parametrization is also a slight improvement over the benzene parametrization for the larger PAHs, without a significant loss of accuracy for the benzene dimer energies, although the description of the crossed benzene configuration is poorer. In particular, in contrast to the W99 potential, the PAHAP potential correctly models the PAH interaction energies at both the stacked as well as the T-shaped configurations. The overall weighted rms residual error over the 238 dimer configurations considered was found to be 0.73 kJ mol^{-1}, which is more than three times less than the error of 2.54 kJ mol^{-1} incurred by the W99 potential.

5. Discussion

Using ab initio calculations, we have developed a transferable anisotropic potential for polycyclic aromatic hydrocarbons that surpasses some of the best empirically derived isotropic potentials in accuracy. In particular, this potential accurately predicts intermolecular interactions for both stacked and non-stacked dimer configurations, such as the T-shape dimer. This is important, while stacked configurations are generally energetically more favorable for most PAH dimers, when modeling clusters of PAHs, as in the context of nascent soot particles, non-stacked configurations are also present.[13,67]

To assess the overall accuracy of our potential, we have to consider the accuracy of both the fit and the SAPT(DFT) energies. The accuracy of the former is shown by the weighted rms residual error, which was calculated to be 0.73 kJ mol^{-1} over the 238 dimer configurations considered. To determine the accuracy of the latter, comparison must be made to other ab initio results. Highly accurate benzene dimer energies have been obtained at CCSD(T) and QCISD(T) levels by Janowski and Pulay.[68] In this work, the largest calculation at QCISD(T)/aug-cc-pVQZ level involved 30 correlated orbitals and 1512 basis functions. Of the three benzene dimers configurations considered, only energy calculations for the T-shape dimer are directly comparable to our SAPT(DFT) calculations. At a separation of 4.989 Å, the QCISD(T) binding energy extrapolated to infinite basis is 11.23 kJ mol^{-1}, whereas the corresponding SAPT(DFT) binding energy calculated at a separation of 5.0 Å in our MC+ basis is 10.33 kJ mol^{-1}. This discrepancy is probably due to the difference in basis sets. The MC+ basis used in our SAPT(DFT) calculations comprises the Sadlej-pVTZ

basis for the monomer centered functions and the extra 3s2p1d midbond functions, but this combined basis is considerably smaller than that used in QCISD(T) calculations and is the likely cause for the underestimation of the binding energy. Using larger basis sets would reduce the error but would increase computational demands prohibitively. Thus, at our chosen level of theory, the error is unavoidable, and given the transferable nature of our PAH potential, we believe this error to be quite acceptable.

The new transferable PAH anisotropic potential represents a first step in our planned investigation of the intermolecular chemistry involved in the clustering of PAHs, which is thought to be an important step in the formation of nascent soot particles. While this potential can be used in its own right, it is hoped that it will also provide an accurate reference against which we can produce a general coarse-grained PAH potential, necessary for the study of large molecular clusters. The potential may also find applications in other fields were the effects of anisotropy could be important, such as organic crystal structure prediction.[69]

Acknowledgment. T.S.T. gratefully acknowledges financial support from the Engineering and Physical Sciences Research Council, Shell Research Ltd. and from Churchill College, Cambridge. The authors thank R. Podeszwa and K. Szalewicz for data provided relating to work reported in ref 18.

Supporting Information Available: Monomer coordinates with partial atomic point charges for benzene, naphthalene, anthracene and pyrene are provided. Axes files which define the local axis system used for each molecule are also provided for use with the Orient program.[19] This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Herdman, J. D.; Miller, J. H. *J. Phys. Chem. A* **2008**, *112*, 6249–6256.

(2) Chen, H. X.; Dobbins, R. A. *Combust. Sci. Technol.* **2000**, *159*, 109–128.

(3) Ishiguro, T.; Takatori, Y.; Akihama, K. *Combust. Flame* **1997**, *108*, 231–234.

(4) Vander Wal, R. L.; Yezerets, A.; Currier, N. W.; Kim, D. H.; Wang, C. M. *Carbon* **2007**, *45*, 70–77.

(5) McKinnon, J. T.; Howard, J. B. *Proc. Combust. Inst.* **1992**, *24*, 965–971.

(6) Miller, J. H. *Proc. Combust. Inst.* **1990**, *23*, 91–98.

(7) Miller, J. H. *Proc. Combust. Inst.* **2005**, *30*, 1381–1388.

(8) Miller, J. H.; Smyth, K. C.; Mallard, W. G. *Proc. Combust. Inst.* **1985**, *20*, 1139–1147.

(9) Schuetz, C. A.; Frenklach, M. *Proc. Combust. Inst.* **2002**, *29*, 2307–2314.

(10) Appel, J.; Bockhorn, H.; Wulkow, M. *Chemosphere* **2001**, *42*, 635–645.

(11) Happold, J.; Grotheer, H.; Aigner, M. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 1247–1254.

(12) Miller, J. H.; Mallard, W. G.; Smyth, K. C. *J. Phys. Chem.* **1984**, *88*, 4963–4910.

(13) Rapacioli, M.; Calvo, F.; Spiegelman, F.; Joblin, C.; Wales, D. J. *J. Phys. Chem. A* **2005**, *109*, 2487–2497.

(14) van de Waal, B. W. *J. Chem. Phys.* **1983**, *79*, 3948–3961.

(15) Williams, D. E. *J. Mol. Struct.* **1999**, *485−486*, 321–347.

(16) Williams, D. E. *J. Comput. Chem.* **2001**, *22*, 1–20.

(17) Williams, D. E. *J. Comput. Chem.* **2001**, *22*, 1154–1166.

(18) Podeszwa, R.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2735–2746.

(19) Stone, A. J.; Dullweber, A.; Engkvist, O.; Fraschini, E.; Hodges, M. P.; Meredith, A. W.; Nutt, D. R.; Popelier, P. L. A.; Wales, D. J. *ORIENT: a program for studying interactions between molecules*, version 4.6; University of Cambridge: Cambridge, U.K., 2002; http://www-stone.ch.cam.ac.uk/programs.html. Accessed on October, 2009.

(20) Weilmünster, P.; Keller, A.; Homann, K. H. *Combust. Flame* **1999**, *116*, 62–83.

(21) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. *J. Chem. Phys.* **2004**, *120*, 647–659.

(22) Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. *Chem. Phys. Lett.* **2000**, *319*, 547–554.

(23) Gerenkamp, M.; Grimme, S. *Chem. Phys. Lett.* **2004**, *392*, 229–235.

(24) Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301–306.

(25) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(26) Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, 214109.

(27) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

(28) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464–470.

(29) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

(30) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

(31) Hesselmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

(32) DiStasio, R. A., Jr.; von Helden, G.; Steele, R. P.; Head-Gordon, M. *Chem. Phys. Lett.* **2007**, *437*, 277–283.

(33) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.

(34) Misquitta, A. J.; Welch, G. W. A.; Stone, A. J.; Price, S. L. *Chem. Phys. Lett.* **2008**, *456*, 105–109.

(35) Podeszwa, R.; Bukowski, R.; Rice, B. M.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5561–5569.

(36) Williams, J. G; Stone, A. J. *J. Chem. Phys.* **2003**, *119*, 4620–4628.

(37) Misquitta, A. J.; Stone, A. J. *J. Chem. Phys.* **2006**, *124*, 024111.

(38) Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2008**, *4*, 7–18.

(39) Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 19–32.

(40) Misquitta, A. J.; Stone, A. J. *Mol. Phys.* **2008**, *106*, 1631–1643.

(41) Stone, A. J.; Misquitta, A. J. *Int. Rev. Phys. Chem.* **2007**, *26*, 193–222.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr. J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*, Gaussian, Inc.: Wallingford, CT, 2004.

(43) Misquitta, A. J.; Stone, A. J. *CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form*, University of Cambridge: Cambridge, U.K., 2007; http://www-stone.ch.cam.ac.uk/programs.html. Accessed on 10/2009.

(44) *DALTON, a molecular electronic structure program, Release 2.0*, 2005, see http://www.kjemi.uio.no/software/dalton/dalton.html. Accessed on 10/2009.

(45) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(46) Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995.

(47) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.

(48) Tozer, D. J. *J. Chem. Phys.* **2000**, *112*, 3507–3515.

(49) Lias, S.; Liebman, J. Ion Energetics Data in NIST Chemistry WebBook, *NIST Standard Reference Database Number 69*, Linstrom P. J. Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, p. 20899; http://webbook.nist.gov. Accessed on October, 2009.

(50) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 7374–7391.

(51) Akin-Ojo, O.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2003**, *119*, 8379–8396.

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

(53) Shoemake, K. Uniform random rotations. In *Graphics Gems III*; Academic Press Professional, Inc.: San Diego, CA, 1992; pp 124−132.

(54) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

(55) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.

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

(57) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

(58) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.

(59) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, 1996; pp 156−184.

(60) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726–3741.

(61) Kim, Y. S.; Kim, S. K.; D., L. W. *Chem. Phys. Lett.* **1981**, *80*, 574.

(62) Misquitta, A. J.; Stone, A. J. Ab initio atom-atom potentials using CamCASP: pyridine as an example, in preparation.

(63) Nobeli, I.; Price, S. L. *J. Phys. Chem. A* **2000**, *103*, 6448–6457.

(64) Hodges, M. P.; Wheatley, R. J. *Chem. Phys. Lett.* **2000**, *326*, 263–268.

(65) Wheatley, R. J. *Mol. Phys.* **1993**, *79*, 597–610.

(66) Bukowski, R.; Sadlej, J.; Jeziorski, B.; Jankowski, P.; Szalewicz, K.; Kucharski, S. A.; Williams, H. L.; Rice, B. M. *J. Chem. Phys.* **1999**, *110*, 3785–3803.

(67) Totton, T. S.; Misquitta, A. J.; Chakrabarti, D.; Wales, D. J.; Kraft, M. Modelling the Internal Structure of Nascent Soot Particles, 2009, Accepted for publication in Combustion and Flame.

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

(69) Price, S. L.; Price, L. S. Modelling intermolecular forces for organic crystal structure prediction. In *Intermolecular Forces and Clusters I: Structure and Bonding*, 2nd ed.; Mingos, D., Wales, D. J., Eds.; Springer-Verlag: Berlin, Heidelberg, Germany, 2005; Vol. 115, pp 81−123.

CT9004883

JCTC Journal of Chemical Theory and Computation

Uncontracted Rys Quadrature Implementation of up to G Functions on Graphical Processing Units

Andrey Asadchev,[†] Veerendra Allada,[‡] Jacob Felder,[†] Brett M. Bode,[‡] Mark S. Gordon,*[,†] and Theresa L. Windus[†]

Department of Chemistry and Department of Electrical and Computer Engineering Iowa State University and Ames Laboratory, Ames, Iowa 50011

Received September 25, 2009

Abstract: An implementation is presented of an uncontracted Rys quadrature algorithm for electron repulsion integrals, including up to **g** functions on graphical processing units (GPUs). The general GPU programming model, the challenges associated with implementing the Rys quadrature on these highly parallel emerging architectures, and a new approach to implementing the quadrature are outlined. The performance of the implementation is evaluated for single and double precision on two different types of GPU devices. The performance obtained is on par with the matrix−vector routine from the CUDA basic linear algebra subroutines (CUBLAS) library.

1. Introduction

The evaluation of two-electron ($2e^-$) repulsion integrals (ERI) is a major computational step in determining the electronic structure of molecules using ab initio quantum chemistry and density functional theory (DFT) methods.[1] Accelerating the integral calculations significantly reduces the overall runtime of the direct Hartree−Fock (HF)[2] and the post-HF methods, e.g., many body perturbation methods.[3]

In 1951, Boys[4] proposed using Gaussian functions as a standard atomic basis set for quantum chemistry computations because the integrals over the Gaussian functions can be evaluated efficiently in closed form. Since then, many different algorithms have been developed to evaluate ERIs over Gaussian functions. The Gauss quadrature method using orthogonal Rys polynomials, developed by Dupuis, Rys, and King (DRK),[5] is a general algorithm that is applicable to a wide range of integrals that arise in computational chemistry. Besides the original Rys quadrature, other ERI algorithms have been developed by, for example, Pople and Hehre (PH),[6] McMurchie and Davidson (MD),[7] Obara and Saika (OS),[8] and Head-Gordon and Pople (HGP).[9] Some of the modifications to the original DRK algorithm are due to Lindh, Ryu, and Liu (LRL)[10] and Dupuis and Marquez (DM).[11] Each of the developed schemes is more efficient for particular cases of integrals, while less efficient or inapplicable for other cases. In practice, quantum chemistry codes, such as the general atomic and molecular electronic structure system (GAMESS),[12] include several different ERI methods in order to take optimal advantage of the best method for particular integral and angular momentum types. The focus of the present work is on ERIs over higher (e.g., d, f) angular momentum functions.

Computationally the ERI calculations scale as $\sim M^3$ to $\sim M^4$, where M is the number of basis functions used in the calculation, and the scalability range depends on the amount of integral prescreening that can be applied. For the most common calculations, M is typically less than a thousand, while larger calculations could require thousands of basis functions.

The ERI calculations are specific to the domain of computational chemistry and the related fields. They are much less common than general methods, such as Fourier transforms and linear algebra kernels, and typically are not as optimized as the basic linear algebra subroutines (BLAS) libraries. A typical HF or DFT calculation requires both ERI and linear algebra computations. However, ERI computations tend to dominate the overall time, since they require more floating point operations (flops). Moreover, unlike numerical linear algebra kernels that exhibit well-defined memory access patterns and simple long loop structures, ERI calculations have to account for many types of integral classes, and therefore, iteration variables do not have a simple linear

* Corresponding author. E-mail: mark@si.msg.chem.iastate.edu.
[†] Department of Chemistry.
[‡] Department of Electrical and Computer Engineering.

10.1021/ct9005079
Published on Web 02/03/2010

relationship to the data elements which must be accessed. To help speed up the time needed to complete the ERI calculations, a general graphical processing unit (GPU) programming model has been implemented. The goal is to implement high angular momentum uncontracted ERIs in this scheme.

2. Electron Repulsion Integrals

Gaussian functions are taken as the standard basis for most ab initio methods. The Cartesian form of a primitive uncontracted one-electron Gaussian basis function with the center located at the origin takes the form of eq 1:

$$\phi(r) = x^{a_x} y^{a_y} z^{a_z} \exp(-\alpha r^2) \tag{1}$$

where α is the Gaussian exponent that governs the spatial extent of the function, r measures the distance from the atomic origin, and a_x, a_y, and a_z are local quantum numbers that determine the net angular momentum L_a by eq 2:

$$L_a = a_x + a_y + a_z \tag{2}$$

Individual Gaussian functions, like those described in eq 1, are generally called "primitive" functions. Especially for lower angular momentum functions (e.g., s and p functions), the actual basis functions are taken to be linear combinations ("contractions") of primitive Gaussians:

$$\phi_a(r) = \sum_k^K D_{ka}\phi_k(r) \tag{3}$$

On the other hand, functions with higher angular momentum (e.g., d, f, and g functions) are typically uncontracted—the focus of this work.

An uncontracted ERI in terms of these one-electron functions can be expressed as:

$$(ab|cd) = \int\int \phi_a(1)\phi_b(1)\frac{1}{r_{12}}\phi_c(2)\phi_d(2)dr_1 dr_2 \tag{4}$$

A contracted ERI can be constructed from a series of uncontracted ERIs (eq 5):

$$(ij|kl) = \sum_a^K \sum_b^L \sum_c^M \sum_d^N D_{ai} D_{bj} D_{ck} D_{dl} (ab|cd) \tag{5}$$

The angular momentum L_a specifies the shape of the function and is denoted by the letters s, p, d, f, etc., for angular momentum values of 0, 1, 2, 3, etc., respectively. Functions with the same angular momentum that differ only in a_x, a_y, and a_z indices belong to the same shell. Grouping functions into shells allows the ERIs to be evaluated more efficiently. The size (i.e., the number of functions) of a shell with angular momentum L_a is

$$\binom{L_a + 2}{2}$$

and the size of an ERI shell block is

$$\binom{L_a + 2}{2}\binom{L_b + 2}{2}\binom{L_c + 2}{2}\binom{L_d + 2}{2}$$

where

$$\binom{n}{2}$$

is the binomial coefficient evaluated as $n(n-1)/2$, and L_a, L_b, L_c, and L_d are the angular momenta of the four atomic orbitals in the ERI.

Note that the individual ERIs have an eight-fold symmetry, since $(ij|kl) = (ji|kl) = (ij|lk)$, etc.; however, the ERIs are computed as shell blocks, rather than individual integrals. So generally, the eight-fold symmetry is only relevant between blocks, not within a block.

2.1. Rys Quadrature. The Rys quadrature proposed by DRK is efficient for higher order integrals (integrals with a higher order angular momentum) that are required for very accurate calculations that include electron correlation. However, it is less efficient for lower order highly contracted integrals. An attractive feature of the Rys quadrature is that it is very stable numerically, an important advantage for higher order integrals. Unlike other methods mentioned in the Introduction Section, it has a very low memory footprint, making it amenable for architectures with smaller caches, such as the GPUs of interest in this work.

The basic idea of the Rys quadrature is to evaluate the integral using a numerical Gaussian quadrature based on a set of orthogonal Rys polynomials. Equation 4 can be expressed, using i, j, k, l to denote functions of a primitive uncontracted ERI shell block $(ab|cd)$, in the form

$$(ij|kl) = \sum_{m=0}^{L} C_m F_m(X) \tag{6}$$

$$F_m(X) = \int_0^1 t^{2m}\exp(-Xt^2)\mathrm{d}t \tag{7}$$

$$\begin{aligned} X &= \rho(r_A - r_B)^2 \\ r_A &= (\alpha_i r_i + \alpha_j r_j)/A \\ r_B &= (\alpha_k r_k + \alpha_l r_l)/B \end{aligned} \tag{8}$$

$$\begin{aligned} \rho &= AB/(A+B) \\ A &= \alpha_i + \alpha_j \\ B &= \alpha_k + \alpha_l \end{aligned} \tag{9}$$

$$L = L_a + L_b + L_c + L_d \tag{10}$$

As suggested in eq 8, X depends on the Gaussian exponents and centers. Equation 6 can be written as eq 11:

$$(ij|kl) = \int_0^1 \exp(-Xt^2)P_L(t)\mathrm{d}t \tag{11}$$

where P_L is a polynomial of degree L, eq 10, with the coefficients C_m in eq 6. Equation 11 can be evaluated exactly by an N-point (where N is an integer greater than $L/2$) Gaussian quadrature:

$$(ij|kl) = \sum_{\omega=1}^{N} W_\omega P_L(t_\omega) \tag{12}$$

$$N = L/2 + 1 \tag{13}$$

The W_α and t_α are weights and roots of the Rys polynomial, respectively. For example, a (*dd*|*dd*) block will have $L = 2 + 2 + 2 + 2 = 8$ and $N = 8/2 + 1 = 5$, and a (*gg*|*ff*) block will have $L = 4 + 4 + 3 + 3 = 14$ and $N = 14/2 + 1 = 8$. Separation of variables allows the terms of the P_L polynomial, which are integrals over dr_1dr_2 (see eq 4) to be written as a product of three two-dimensional (2-D) integrals I_x, I_y, and I_z over dx_1dx_2, dy_1dy_2, and dz_1dz_2, respectively.

$$P_L(t) = 2(\rho/\pi)^{1/2} I_x I_y I_z \tag{14}$$

The overall ERI formula becomes

$$(ij|kl) = 2(\rho/\pi)^{1/2} \sum_\omega I_x(t_\omega) I_y(t_\omega) I_z(t_\omega) W_\omega \tag{15}$$

The 2-D integrals of a shell block have array dimensions (in FORTRAN/MATLAB notation, where commas delimit dimensions and colons specify the range) as shown below:

$$I_{q(=x,y,z)}(N, 0{:}L_a, 0{:}L_b, 0{:}L_c, 0{:}L_d) \tag{16}$$

The first or leading dimension in eq 16 corresponds to the number of roots, and the last four dimensions correspond to Cartesian exponents for each function in a shell block. When constructing the ERI block, the 2-D integrals will be reused multiple times, hence, the computational and memory advantage of calculating ERIs as a block. For example, to construct the first six integrals of the (*pp*|*pp*) shell block, the following 2-D Cartesian integrals are used (multiplication by a constant factor is implied):

$$\begin{aligned}
(p_x p_x|p_x p_x) &= \sum_\omega I_x(\omega,1,1,1,1) I_y(\omega,0,0,0,0) I_z(\omega,0,0,0,0)\\
(p_y p_x|p_x p_x) &= \sum_\omega I_x(\omega,0,1,1,1) I_y(\omega,1,0,0,0) I_z(\omega,0,0,0,0)\\
(p_z p_x|p_x p_x) &= \sum_\omega I_x(\omega,0,1,1,1) I_y(\omega,0,0,0,0) I_z(\omega,1,0,0,0)\\
(p_x p_y|p_x p_x) &= \sum_\omega I_x(\omega,1,0,1,1) I_y(\omega,0,1,0,0) I_z(\omega,0,0,0,0)\\
(p_x p_y|p_x p_x) &= \sum_\omega I_x(\omega,0,0,1,1) I_y(\omega,1,1,0,0) I_z(\omega,0,0,0,0)\\
(p_y p_y|p_x p_x) &= \sum_\omega I_x(\omega,0,0,1,1) I_y(\omega,0,1,0,0) I_z(\omega,1,0,0,0)
\end{aligned}$$

The roots and weights of the Rys polynomials can be evaluated by polynomial approximations[13] or by using a general Stieltjes procedure.[14] The 2-D Cartesian integrals are evaluated efficiently using recurrence and transfer relationships. The recurrence relationships generate 2-D integrals with all angular momenta shifted to centers *i* and *k* from (*ss*|*ss*) 2-D integrals, and transfer relationships shift the angular momentum to centers *j* and *l* to generate the desired 2-D integrals. For the details of these relationships, the reader is referred to the original DRK paper.[5]

The quadrature step itself, i.e., the summation over the roots, eq 15, is the time-consuming step of the ERI shell calculation, requiring:

$$3N\binom{L_a+2}{2}\binom{L_b+2}{2}\binom{L_c+2}{2}\binom{L_d+2}{2}$$

flops, where the factor of 3 is from the two multiplications and an addition in each step. The transfer relationships scale as $N(L_a+1)(L_b+1)(L_c+1)(L_d+1)$, requiring many fewer operations than the quadrature step as the angular momentum increases. The recurrence relationships and root evaluation require even fewer flops than the transfer relationships for higher order integrals. Therefore, since an efficient implementation of the quadrature step determines the overall performance of the algorithm, the main topic of this paper is the efficient parallel implementation of the quadrature. Unlike the recurrence and transfer equations, which have predictable memory access patterns and can be expressed as simple vector operations, the quadrature step has complex memory access patterns which span a large data set and depend on the particular ERI class being evaluated. For example, the evaluation of the (*ff*|*ff*) ERI block requires $3N(L_a+1)(L_b+1)(L_c+1)(L_d+1) = (3)(7)(4)(4)(4)(4) = 5376$ floating point (FP) numbers for three 2-D integral arrays (*X*, *Y*, *Z*) and $10^4 = 10000$ FP numbers for the final integral. For double precision numbers, the overall memory would be 123008 Bytes, well beyond the size of a typical L1 data cache.

Algorithm 1 outlines the basic structure of the Rys quadrature. Its simplicity obscures the fact that the Cartesian indices do not have a simple relationship to the iteration variables and must be either tabulated or each case must be programmed specifically for a particular ERI class.

Algorithm 1: Rys Quadrature

Algorithm 1 Rys Quadrature

```
for all l do
  for all k do
    for all j do
      for all i do
        I(i,j,k,l) = Σ_ω I_x(ω,i_x,j_x,k_x,l_x) I_y(ω,i_y,j_y,k_y,l_y) I_z(ω,i_z,j_z,k_z,l_z)
      end for
    end for
  end for
end for
```

3. Graphical Processing Units

GPU technology has emerged as a viable computing platform for general purpose application programming, also known as general purpose computation on graphical processing units (GPGPU). The GPUs offer high-density arithmetic units at the expense of larger cache sizes and control units. In terms of linear algebra kernels, the GPUs can approach 20 and 70 giga floating point operations per second (GFLOPS) for matrix−vector and matrix−matrix routines, respectively, on current double precision (DP) capable devices[15] that have a theoretical peak of around 90 GFLOPS.

3.1. Compute Unified Device Architecture. Among the current GPGPU technologies, the NVIDIA compute unified device architecture (CUDA)[16] language environment is available for several GPU devices and is the target implementation choice. CUDA is a unified hardware computing architecture and programming model for graphics as well

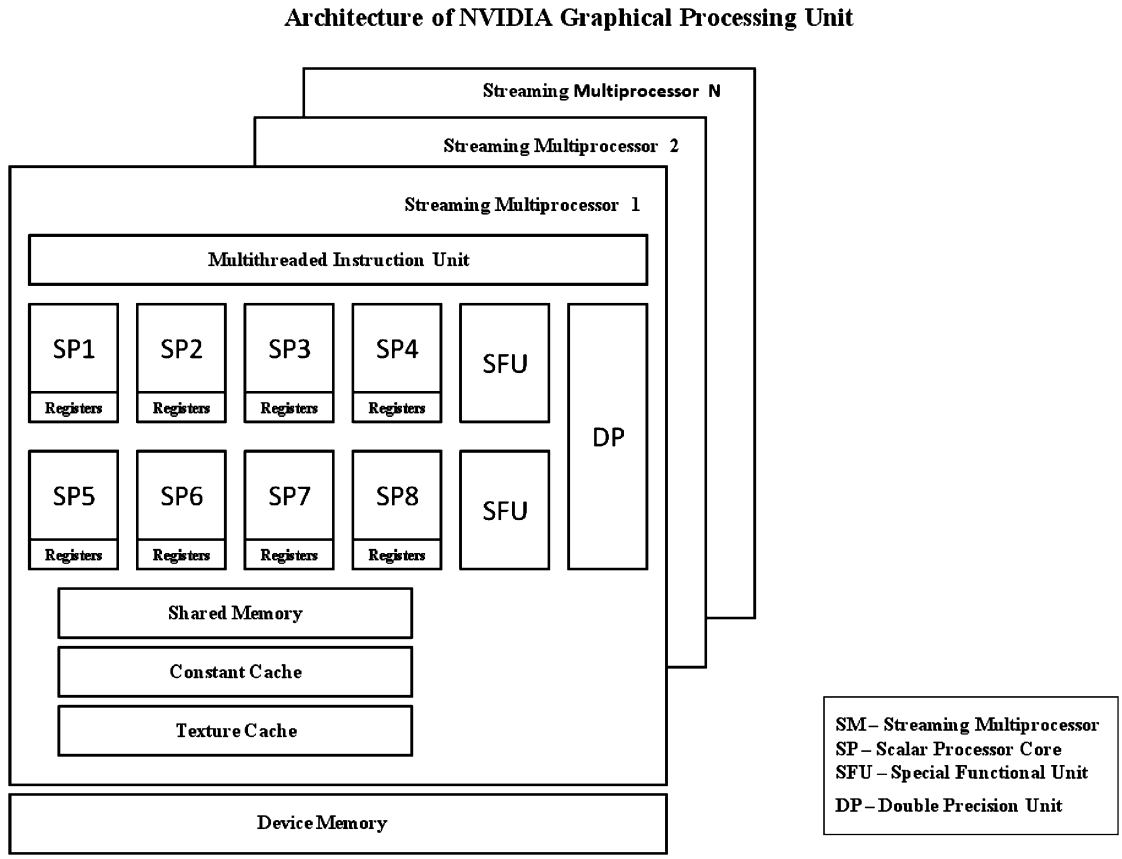


Figure 1. High-level architecture of a GPU.

as general-purpose processors. The current CUDA device architecture consists of a scalable array of streaming multiprocessors (SM), see Figure 1. Each SM consists of eight scalar processors (SP), a multithreaded instruction unit, on-chip shared memory, one double precision unit, and two special purpose transcendental functional units. Under the CUDA programming model, the GPU is viewed as a highly multithreaded compute device capable of executing many threads in parallel. The threads execute a sequence of instructions in a data parallel fashion—single-instruction multiple threads (SIMT).

Computationally demanding code paths of an application are isolated into functions (*kernels* in NVIDIA terminology) that are compiled into the instruction set architecture of the GPU device. The CUDA programming interface is designed with a minimal set of extensions to the C/C++ language. A runtime library provides functions to manage the compute device, to perform memory operations, and to run the device-specific functions. The main goal of the programming environment is to develop scalable and efficient parallel programs.

A computational kernel is launched from the host and executed by T threads (T is application specific) on the device. The threads are hierarchically arranged as a grid of blocks and as a block of threads, as shown in Figure 2 (adopted from the programmers manual).[17] Each thread within a thread block has a unique set of (x, y, and z) indices that allow three-dimensional (3-D) data to be mapped onto

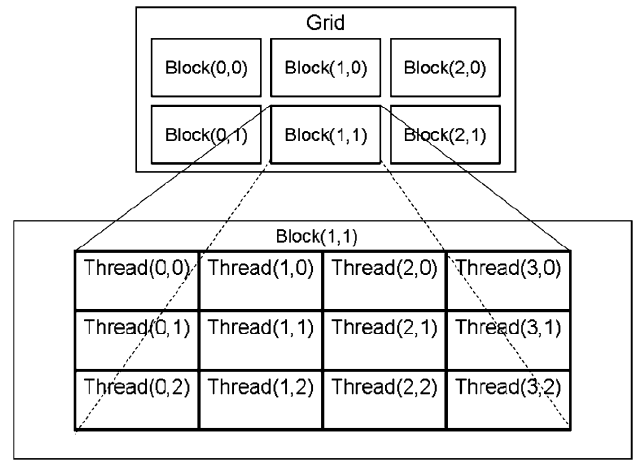


Figure 2. Grid of blocks and block of threads (z-dimension is implied).

a thread block. Each thread block has unique x and y coordinates, which map all the thread blocks onto a 2-D grid of blocks.

The logical memory space seen by the threads can be hierarchically arranged based on the data visibility (see Figure 1). Each thread has access to its local registers on the processor. Threads in a block can access and share data via the parallel data cache, called shared memory. The registers and the shared memory have a low latency and are limited resources available to the threads. One of the biggest challenges in designing the kernels lies in optimizing the per-thread register and the shared memory usage. Each thread also has access to a private local memory and a global

memory space that are both part of the device memory and have high data access latencies. Apart from these, an application can also use the read-only constant and the texture memories that are cached.

Access to the main memory has a high latency, on the order of hundreds of cycles. To achieve full bandwidth, accesses to the main memory must be coalesced, meaning consecutive threads access consecutive memory elements to achieve full memory bandwidth. Coalescing ensures that multiple memory requests are being served simultaneously rather than served sequentially. Although only one thread block can execute at any given time on a SM, multiple thread blocks can be *active*, thereby hiding the memory latency by overlapping the computations and the communications. An active block in CUDA terminology is a block that is ready for execution whenever a SM becomes free, e.g., when the current executing thread block starts fetching the memory. The number of active thread blocks is limited by the register and the shared memory usage.

The execution of a thread block is further batched into a series of *warps* that are consecutively arranged based on the thread number in batches of 32. To maximize the parallel performance, all threads in a warp must execute an identical GPU instruction, otherwise the warp is said to diverge, and the differing instructions are executed sequentially.

4. CUDA Rys Quadrature Implementation

4.1. Related Work. Ufimtsev and Martinez[18] evaluated the ERIs of *s* and *p* functions on GPUs in single precision using the MD algorithm. They later developed an entire Hartree−Fock code that runs on a GPU and showed improved performance[19] over the CPU code. Yasuda[20] implemented the Rys quadrature on a GPU in enhanced single precision for *s* and *p* integrals (i.e., some double precision computations were emulated in the software but still using single precision hardware). A new interpolation formula was proposed for the roots and the weights, and an error analysis for the quadrature was given. Some work has also been done to implement ERI algorithms on IBM CELL and FPGAs,[21] however only for the rather limited case of (*ss*|*ss*) ERIs.

To our knowledge there has not been a reported implementation of an ERI algorithm for *d* or higher angular momentum functions on GPUs or on other accelerators. The main difficulty seems to have been the limited amount of fast memory and the amount of code that must be generated for many cases involving higher angular momentum functions. This is the focus of the present work.

4.2. Implementation Considerations. Since the ERI computations are memory bound, the main consideration in designing the CUDA Rys quadrature is to optimize the memory access patterns and the data reuse. The 2-D integrals are reused multiple times to construct different ERIs and should, therefore, be loaded into shared memory. This also implies that an ERI block should be mapped onto a single thread block, as shared memory access and synchronization is limited to thread block boundaries. The ERI blocks are mapped onto the grid so that each thread block computes one ERI block. For the purposes of discussion, block is used to refer to both thread and ERI blocks.

Device memory loads and stores should be coalesced to parallelize memory accesses with high latencies. Multiple thread blocks should be active on a single SM in order to hide memory latency by overlapping computation and communication. In order to have multiple active thread blocks, the shared memory and the registers should be used sparingly. To illustrate hardware constraints, if a GPU has only 1638 4-byte registers and 16 KB of shared memory, then a kernel using 32 registers per thread and 2688 bytes of shared memory per thread block has a limit of 512 threads imposed by the register use and 6 active thread blocks per SM due to shared memory availability.

Clearly, for the larger ERI classes, the entire set of 2-D integrals cannot be kept in shared memory all at once but must be loaded from the device memory as needed. The iteration through the ERIs should be done so as to minimize the number of device memory loads. If the ERIs are only computed on the GPU but are not contracted right away, e.g., to form the Fock operator, then there is no reuse of the final ERIs.

4.3. Implementation Design. The current CUDA capable hardware imposes a limit of 512 or 768 maximum threads per block, depending on the particular GPU device. Consider the (*dd*|*dd*) ERI block case. The size of the entire block is $6^4 = 1296$ elements, exceeding the maximum number of threads. However, it is possible to map multiple elements to a single thread, e.g., by mapping *i*, *j*, and *k* indices, corresponding to the first three shells of the block, to a unique thread and iterating over the last index *l*. Since the thread block is 3-D, the mapping of the *i*, *j*, and *k* shell index to a thread is natural. Algorithm 2 outlines the general idea. The algorithm is in Python-like pseudo code, with ## signifying comments, and the indices and loops over the roots, *N*, are implied.

Algorithm 2: CUDA Rys Quadrature, *i, j,* and *k* Mapping

Algorithm 2 CUDA Rys quadrature, i, j, k mapping

```
## map threads to ERI elements threadIdx is the thread coordinate
i = threadIdx.x
j = threadIdx.y
k = threadIdx.z
## The arrays LX, LY, LZ map functions to exponents
(ix, iy, iz) = (LX[i], LY[i], LZ[i])
(jx, jy, jz) = (LX[j], LY[j], LZ[j])
(kx, ky, kz) = (LX[k], LY[k], LZ[k])
for all / do
  sync threads
  ## load 2-D integrals to shmem
  if LX[l] ≠ LX[l − 1] then
    I_x,shmem = I_x(:, :, :, LX[l])
  end if
  if LY[l] ≠ LY[l − 1] then
    I_y,shmem = I_y(:, :, :, LY[l])
  end if
  if LZ[l] ≠ LZ[l − 1] then
    I_z,shmem = I_z(:, :, :, LZ[l])
  end if
  sync threads
  I(i, j, k, l) = ∑_N I_x,shmem(ix, jx, kx) I_y,shmem(iy, jy, ky) I_z,shmem(iz, jz, kz)
end for
```

In terms of shared memory, i, j, and k mapping requires all the 2-D integrals of a specific l index. For the (dd|dd) ERI case, this means that the shared memory overhead for each l iteration is $N(L_a + 1)(L_b + 1)(L_c + 1) = 5(3^3) = 135$ elements per 2-D integral block. Though three 2-D integral blocks are needed per iteration, it is most likely that one of the previous Cartesian indices will stay the same. This means that the corresponding 2-D integral is already in the shared memory, reducing the memory communication by a third. For example, the construction of a d shell is outlined below. The three rows correspond to I_x, I_y, and I_z Cartesian indices. The indices marked with an asterisk represent load operations. Though there are a total of 18 indices, only 13 indices must be loaded if the shell is arranged to minimize loads.

$$\begin{matrix} I_x \\ I_y \\ I_z \end{matrix}\begin{pmatrix} 0^* \\ 2^* \\ 0^* \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 0^* \\ 2^* \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 1^* \\ 1^* \end{pmatrix} \rightarrow \begin{pmatrix} 1^* \\ 1 \\ 0^* \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0^* \\ 1^* \end{pmatrix} \rightarrow \begin{pmatrix} 2^* \\ 0 \\ 0^* \end{pmatrix}$$

The above order may differ from the requirements of an application, however, restoring the desired ordering is trivial. In the (*ff*|*ff*) ERI case, mapping three indices to threads is not possible, as it requires 1000 threads. However, we can map the i and j indices and loop over the k and l indices in a similar fashion, as outlined in Algorithm 3.

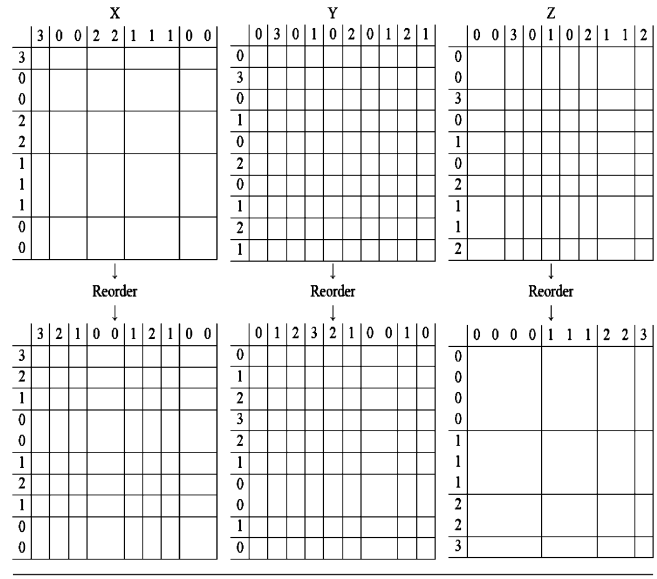
Algorithm 3: CUDA Rys Quadrature, i and j Mapping

Algorithm 3 CUDA Rys quadrature, i, j mapping

```
## map threads to ERI elements
i = threadIdx.x
j = threadIdx.y
## lookup Cartesian exponents
(ix, iy, iz) = (LX[i], LY[i], LZ[i])
(jx, jy, jz) = (LX[j], LY[j], LZ[j])
for all kl_{z-block} do
  sync threads
  I_{z,shmem} = I_z(:, :, LZ[k], LZ[l])
  ## load 2-D integrals to shmem
  for all kl_{xy} ∈ kl_{z-block} do
    sync threads
    I_{x,shmem} = I_x(:, :, LX[k], LX[l])
    I_{y,shmem} = I_y(:, :, LY[k], LY[l])
    sync threads
    I(i, j, k, l) = Σ_N I_{x,shmem}(ix, jx) I_{y,shmem}(iy, jy) I_{z,shmem}(iz, jz)
  end for
end for
```

The shared memory requirement for an (*ff*|*ff*) ERI is $N(L_a + 1)(L_b + 1) = 7(4^2) = 112$ elements for each 2-D memory block. The memory access can likewise be reduced by a third, if the shells are reordered. Blocking of kl_z indices was used, as outlined in Table 1 for a |*ff*) case. In the example, the number of memory loads is 216. The first row of Table 1 shows the data access pattern for x, y, and z 2-D integrals when the canonical ERI ordering is used. There are some block patterns that are visible in the x and z dimensions. But these blocking patterns are not optimal from the perspective of data reuse because the blocks are small. To improve the overall memory performance, the integrals can be reordered such that one of the 2-D integrals has a well-defined block structure, for example, the z integral.

Table 1. Index Ordering for |*ff*) Case



4.4. Template-based Code Generation. If the cases described above are implemented in CUDA directly, then the register usage is high. To reduce the register use, the loops over the outer indices can be unrolled explicitly for each possible case, e.g. for |*pp*), |*pd*), |*pf*), etc. Programming all of the cases by hand is prohibitive, as it requires a large amount of code. However, using a template-based approach, all of the cases can be generated automatically from a single template.

There exists a number of template engines, e.g., the venerable *m*4 macro processor,[22] but the Python-based[21] Cheetah template engine[24] is chosen for this project. In Cheetah templates, the Python statements that control the code generation are embedded directly in the source code, similar to the manner in which traditional C preprocessor directives are used. Other benefits of using Cheetah are the ability to write complex support modules in Python and to reuse existing Python utilities.

Since generating code from a template is straightforward, the root summation loops were also explicitly unrolled. This was done to allocate registers to store a single set of 2-D integrals in registers rather than in shared memory. The benefit of doing so is that the use of shared memory and the bank conflicts are reduced. All of the shared memory is arranged in banks; the number of banks for the current hardware is 16, i.e., half-warp size. A bank conflict arises when multiple threads in a half-warp access different memory locations mapped onto the same bank, simultaneously resulting in the serialization of threads in the half-warp.[17] The bank conflicts occur often if the leading dimension (in this case, the number of roots, N) is a divisor of the bank size; accesses to an array with a leading dimension of 8 causes 8 bank conflicts. For other cases, the bank conflicts occur much less often; a leading dimension of 7 causes only one bank conflict. The bank conflicts lead to warp serialization, where the warp threads execute the instructions sequentially rather than executing the same instruction in a single instruction multiple thread (SIMT) or a lock-step fashion. Warp serialization is highly undesirable, and bank

conflicts are a serious performance issue, degrading the overall performance by about 25%. The bank conflicts are especially pronounced when the number of roots is even.

5. Results and Discussion

The performance of the implementation is evaluated on the NVIDIA GeForce GTX 275 and the Tesla T10 processors. The GTX 275 is a regular graphics card with a double precision support, an 1 GB of device memory, and a clock speed of 1.1 GHz. The Tesla T10 is a dedicated HPC accelerator with double precision (DP) support, 4 GB of device memory, and a clock speed of 1.35 GHz. The Tesla processor is capable of delivering approximately 20 GFLOPS on the level 2 BLAS double precision general matrix vector (DGEMV) routine and 70 GFLOPS on the level 3 BLAS double precision general matrix multiply (DGEMM) routine. The GTX 275 is approximately 25% slower than the Tesla.

The basis set used was purely synthetic with exponents of 1.5, which is representative of the values for higher angular momentum functions. The quoted timings do not include the GPU−CPU communication time.

The performance of the quadrature was evaluated by counting the total number of quadrature operations given by

$$\text{flops} = n_{\text{block}} 3N \binom{L_a+2}{2}\binom{L_b+2}{2}\binom{L_c+2}{2}\binom{L_d+2}{2}$$

where n_{block} is the number of ERI/thread blocks. The above metric also accounts for multiplication by a constant factor that incorporates contraction coefficients and normalization factors. For example, the (*ff*|*ff*) block requires a $21(10^4) = 210000$ flop count. The total flop count is divided by the execution time on the GPU to obtain the GFLOPS metric. The execution time does not include the memory transfer overheads between the host and the GPU. The transfer time latency of the ERIs from GPU to host is several times longer than that of the actual execution time.

The performance results on the GTX 275 and Tesla boards are presented in Tables 2 and 3, respectively. As can be seen from these tables, the performance depends to a large degree on the ERI class. The larger ERI classes (i.e., higher collective angular momentum) perform better on average than the smaller classes. The computations with an odd number of roots, cf., eq 13, e.g., (*gg*|*dd*), (*ff*|*ff*), etc., tend to have fewer bank conflicts than those with an even number of roots, e.g., (*gg*|*ff*), as discussed in Section 4. Consequently, the performance of ERI classes with an odd number of roots is higher, as much as 25% in an extreme case. The difference between the single and double precision performance is roughly a factor of 2, as previously predicted by Ufimtsev and Martinez.[25] One would expect this difference to favor single precision even more strongly, since the number of SP units is eight times the number of DP units. This suggests that the computations are memory bound rather than compute bound. The performance depends heavily on the mapping used, cf., Section 4.3. As one would expect, the "larger" *i*, *j*, and *k* mapping performs better than the *i* and *j* mapping for cases with lower *i* and *j* angular momenta (such as the (*pp*|*ff*) ERI block), since the shared memory reuse and parallelism is much higher. The difference between the two mappings for the same ERI class can be as high as a factor of 5. However, when the *i* and *j* angular momenta are higher (such as the (*ff*|*pp*) ERI block), the *i* and *j* mapping is only slightly outperformed by the *i*, *j*, and *k* mapping. Interestingly, comparing the best performance for the (*pp*|*ff*) and (*ff*|*pp*) ERI block gives very similar performance. One could use ERI index symmetry such as (*ij*|*kl*) = (*kl*|*ij*) in these cases to ensure that the first two indices are always lower, so the *i*, *j*, and *k* mapping algorithm could always be used when the memory is available.

The difference in performance between the generic GTX GPU and the Tesla T10, presented in Table 4, is 25−30% across the single and double precision performance. This is

Table 2. CUDA Rys Quadrature Performance on GeForce GTX 275

			GFLOPS$_{SP}$[c]		GFLOPS$_{DP}$[d]	
ERI	blocks[a]	flop count[b]	map$_{ijk}$[e]	map$_{ij}$[e]	map$_{ijk}$[e]	map$_{ij}$[e]
(*gg*|*gg*)	2000	2733750000	n/a	**45.23**	n/a	**22.55**
(*gg*|*ff*)	4000	2160000000	n/a	**34.42**	n/a	**15.32**
(*ff*|*gg*)	4000	2160000000	n/a	**30.91**	n/a	**14.11**
(*gg*|*dd*)	10000	1701000000	n/a	**43.08**	n/a	**21.05**
(*dd*|*gg*)	10000	1701000000	n/a	**23.63**	n/a	**16.35**
(*gg*|*pp*)	40000	1458000000	n/a	**36.53**	n/a	**17.08**
(*pp*|*gg*)	40000	1458000000	**34.23**	6.93	**18.20**	5.38
(*ff*|*ff*)	10000	2100000000	n/a	**40.43**	n/a	**20.11**
(*ff*|*dd*)	20000	1296000000	n/a	**37.54**	n/a	**18.29**
(*dd*|*ff*)	20000	1296000000	**37.69**	23.32	**16.53**	15.04
(*ff*|*pp*)	80000	1080000000	27.43	**31.46**	15.23	**17.05**
(*pp*|*ff*)	80000	1080000000	**32.23**	6.21	**17.45**	4.84
(*dd*|*dd*)	60000	1166400000	**31.10**	20.17	**16.38**	13.67
(*dd*|*pp*)	200000	777600000	19.71	**20.25**	11.54	**11.70**
(*pp*|*dd*)	200000	777600000	**20.18**	5.16	**11.11**	3.85
(*pp*|*pp*)	750000	546750000	**11.93**	4.79	**8.43**	3.76

[a] Blocks are the number of ERI blocks evaluated. [b] Flop count is the total floating point operations. [c] GFLOPS$_{SP}$ is the single precision performance. [d] GFLOPS$_{DP}$ is the double precision performance. [e] Map is the ERI to thread mapping; the best performing mapping is shown in bold.

Table 3. CUDA Rys Quadrature Performance on Tesla GPU

			GFLOPS$_{SP}$[c]		GFLOPS$_{DP}$[d]	
ERI	blocks[a]	flop count[b]	map$_{ijk}$[e]	map$_{ij}$[e]	map$_{ijk}$[e]	map$_{ij}$[e]
(*gg*|*gg*)	2000	2733750000	n/a	**55.97**	n/a	**27.34**
(*gg*|*ff*)	4000	2160000000	n/a	**42.07**	n/a	**18.67**
(*ff*|*gg*)	4000	2160000000	n/a	**37.70**	n/a	**17.19**
(*gg*|*dd*)	10000	1701000000	n/a	**53.39**	n/a	**25.34**
(*dd*|*gg*)	10000	1701000000	n/a	**31.71**	n/a	**19.87**
(*gg*|*pp*)	40000	1458000000	n/a	**45.15**	n/a	**20.65**
(*pp*|*gg*)	40000	1458000000	**42.42**	7.78	**22.09**	6.19
(*ff*|*ff*)	10000	2100000000	n/a	**50.19**	n/a	**24.46**
(*ff*|*dd*)	20000	1296000000	n/a	**46.15**	n/a	**22.44**
(*dd*|*ff*)	20000	1296000000	**45.71**	28.46	**19.71**	18.29
(*ff*|*pp*)	80000	1080000000	33.86	**39.38**	18.54	**20.10**
(*pp*|*ff*)	80000	1080000000	**40.33**	7.02	**21.46**	5.63
(*dd*|*dd*)	60000	1166400000	**38.74**	23.38	**19.78**	15.62
(*dd*|*pp*)	200000	777600000	24.67	**25.00**	14.20	**14.33**
(*pp*|*dd*)	200000	777600000	**25.22**	7.67	**13.73**	4.33
(*pp*|*pp*)	750000	546750000	**14.17**	5.37	**10.00**	4.30

[a] Blocks are the number of ERI blocks evaluated. [b] Flop count is the total floating point operations. [c] GFLOPS$_{SP}$ is the single precision performance. [d] GFLOPS$_{DP}$ is the double precision performance. [e] Map is the ERI to thread mapping; the best performing mapping is shown in bold.

Table 4. GTX 275, Tesla, and GAMESS Performance Comparison

			$GFLOPS_{SP}$[c]		$GFLOPS_{DP}$[d]		GLFOPS[e]
ERI	blocks[a]	flop count[b]	GTX 275	Tesla	GTX 275	Tesla	GAMESS
(*gg*|*gg*)	2000	2733750000	45.23	55.97	22.55	27.34	1.36
(*gg*|*ff*)	4000	2160000000	34.42	42.07	15.32	18.67	1.29
(*ff*|*gg*)	4000	2160000000	30.91	37.70	14.11	17.19	1.32
(*gg*|*dd*)	10000	1701000000	43.08	53.39	21.05	25.34	1.09
(*dd*|*gg*)	10000	1701000000	23.63	24.03	16.35	29.88	1.21
(*gg*|*pp*)	40000	1458000000	36.53	45.15	17.08	20.65	0.82
(*pp*|*gg*)	40000	1458000000	34.23	42.42	18.20	22.09	0.98
(*ff*|*ff*)	10000	2100000000	40.43	50.19	20.11	24.46	1.19
(*ff*|*dd*)	20000	1296000000	37.54	46.15	18.29	22.44	0.94
(*dd*|*ff*)	20000	1296000000	37.69	45.71	16.53	19.71	1.03
(*ff*|*pp*)	80000	1080000000	31.46	39.38	17.05	20.10	0.75
(*pp*|*ff*)	80000	1080000000	32.23	40.33	17.45	21.46	0.78
(*dd*|*dd*)	60000	1166400000	31.10	38.74	16.38	19.78	0.79
(*dd*|*pp*)	200000	777600000	20.25	25.00	11.70	14.33	0.63
(*pp*|*dd*)	200000	777600000	20.18	25.22	11.11	13.73	0.66
(*pp*|*pp*)	750000	546750000	11.93	14.17	8.43	10.00	0.48

[a] Blocks are the number of ERI blocks evaluated. [b] Flop count is the total floating point operations. [c] $GFLOPS_{SP}$ is the single precision performance. [d] $GFLOPS_{DP}$ is the double precision performance. [e] Map is the ERI to thread mapping; the best performing mapping is shown in bold.

consistent with the higher clock speed of the Tesla compute device. In terms of registers and shared memory, Tesla does not have an advantage over the generic GTX GPU. This is reflected in the same mapping on both GPUs having the best performance for a particular ERI case. The performance of the Rys quadrature is on par with the performance of the (S/G) DGEMV routines in the CUBLAS library. In terms of the peak theoretical performance, it is possible to achieve approximately 30% in the best case. The poor-performing lower angular momentum ERI classes utilize the hardware at 10% efficiency in the worst case. However, the two mapping implementations (*ijk* and *ij*) are not specifically optimized for these ERI classes.

Previous work in this field focused on *s* and *p* integrals, which produce small, highly contracted integral blocks that can be performed entirely in shared memory and registers. Therefore, the ratio of computation to memory traffic is high. Moreover, on Telsa and older architectures, there is a ratio of 8:1 for single vs double precision floating point units. For higher angular momentum integrals, computation cannot be done entirely in shared memory and registers, so partial values must be read and stored in global memory. This helps to explain the difference in performance between this work and previous work.

The difference in performance between the GPU and the CPU, which is also presented in Table 4, is very promising. The Rys quadrature used in GAMESS, which was also used as a benchmark by Ufimtsev and Martinez, is a legacy FORTRAN implementation that underperforms on modern CPUs. As can be seen from Table 4, the original Rys quadrature implementation is only 15% efficient at best on a modern 8 GFLOP processor. In order to achieve good performance on both CPU and GPU, the algorithm must be implemented in a way suitable for instruction level parallelism.

6. Conclusions and Future Work

This work has demonstrated the ability to obtain comparable or better performance to that of an optimized DGEMV routine for a Rys quadrature implementation of two-electron integral computations—a core computation for electronic structure algorithms. Since the focus of this work is on higher angular momentum integrals, the use of the double precision units on the GPU is also highlighted. In order to achieve the best performance, memory access patterns and data reuse have been optimized. The code implementation has been greatly facilitated by using the template-based code generator. Not only does it allow for fast prototyping of various algorithms, it also provides a developer-friendly framework for the developer to focus on the main issues associated with the algorithm and allows the details associated with the many angular momenta and single vs double precision cases to be handled automatically. The use of templates could eventually be taken one step further, so that the generated code could be optimized further, depending on the GPU architecture. However, this may prove to be impractical as the NVIDIA compiler is refined to take advantage of different GPU architectures.

Some improvements are still possible with respect to data reuse, but the gains are unlikely to be high. The improvements would be due to more aggressive memory caching and memory access pattern reordering. The greatest overall improvement will come from reusing the ERI blocks as soon as they are formed on the GPU, e.g., to construct the Fock matrix. The 2-D Fock matrix is formed from 4-D ERI blocks; so, if computed on the GPU device, then the memory transfer would just be that of the Fock matrix (of order M^2 where M is the number of basis functions) rather than those of all the ERI blocks. Therefore, the computation of the Fock matrix on the GPU increases the flop count and reduces the amount of memory to be transferred to the host, resulting in overall greater performance.[17] The direct use of the ERIs on the GPU device is necessary, as the memory transfer of the raw ERIs between GPU and host is several times longer than the computation itself. The amount of data that must be transferred from the host to the GPU to start the computation is small; moreover, since it is small, it can be transferred asynchronously while the computation is running.

The contracted ERIs and the ERIs of small angular momentum functions have not been addressed directly in this work. The implementation of the Rys quadrature roots recurrence and transfer relationships is also not discussed explicitly but will be presented in a future publication. The accuracy and utility of single precision vs double precision computations will be considered in future work. In addition, future work will include the incorporation of the ERIs into modern algorithms for full electronic structure theory calculations.

Acknowledgment. This work has been accomplished by Iowa State University of Science and Technology under contract no. DE-AC02-07CH11358 with the United States Department of Energy and under an National Science Foundation PetaApps grant. The GPU cluster was provided by a DURIP grant from the Department of Defense, with matching funds from the Iowa State University Frances M. Craig Professorship to M.S.G. and the NVIDIA Corporation. M.S.G. gratefully acknowledges an IBM Faculty Fellowship award. The authors have benefitted from many helpful discussions with Professor Todd Martinez and his students.

References

(1) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(2) Almlöf, J.; Faegri, K., Jr.; Korsell, K. *J. Comput. Chem.* **1982**, *3*, 385–399.

(3) Bartlett, R. J.; Purvis, G. D. *Int. J. Quantum Chem.* **1978**, *14*, 561–581.

(4) Boys, S. F. *Proc. R. Soc. Lond. A* **1950**, *200*, 542.

(5) Rys, J.; Dupuis, M.; King, H. *J. Comput. Chem.* **1983**, *4*, 154–157.

(6) Pople, J.; Hehre, W. *J. Comput. Phys.* **1978**, *27*, 161–168.

(7) McMurchie, L. E.; Davidson, E. R. *J. Comput Phys.* **1978**, *26*, 218–231.

(8) Obara, S.; Saika, A. *J. Chem. Phys.* **1988**, *89*, 1540–1559.

(9) Head-Gordon, M.; Pople, J. A. *J. Chem. Phys.* **1988**, *89*, 5777–5786.

(10) Lindh, R.; Ryu, U.; Liu, B. *J. Chem. Phys.* **1991**, *95*, 5889–5897.

(11) Dupuis, M.; Marquez, A. *J. Chem. Phys.* **2001**, *114*, 2067–2078.

(12) Gordon, M. S.; Schmidt, M. W. Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry: the First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005.

(13) Rys, J.; Dupuis, M.; King, H. *J. Comput. Phys.* **1976**, *21*, 144.

(14) Sagar, R. P.; Smith, V. H., Jr. On the calculation of Rys polynomials and quadratures. *Int. J. Quantum Chem.* **1992**, *42*, 827–836.

(15) Volkov, V.; Demmel, J. W. Benchmarking GPUs to tune dense linear algebra. In *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*; IEEE Press: Piscataway, NJ, 2008.

(16) Nickolls, J.; Buck, I.; Garland, M.; Skadron, K. *ACM Queue* **2008**, *6*, 40–53.

(17) *NVIDIA CUDA Programming Guide*; NVIDIA Corporation: Santa Clara, CA, 2009; http://developer.download.nvidia.com/compute/cuda/2_3/toolkit/docs/NVIDIA_CUDA_Programming_Guide_2.3.pdf.

(18) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.

(19) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.

(20) Yasuda, K. *J. Comput. Chem.* **2008**, *29*, 334–342.

(21) Shi, G.; Kindratenko, V.; Ufimtsev, I. S.; Martinez, T. J. Two-Electron Integral Evaluation on FPGA, CELL and GPU accelerators. Poster presentation in Path to Petascale: *Adapting GEO/CHEM/ASTRO Applications for Accelerators and Accelerator Clusters Workshop*; National Center for Supercomputing Applications: Urbana, IL, 2009.

(22) *GNU m4 Macro Processor Manual*; Free Software Foundation, Inc.: Boston, MA, 2007; http://www.gnu.org/software/m4/manual/m4.html.

(23) Pilgrim, M. *Dive Into Python*; CreateSpace: Paramount, CA, 2009.

(24) Martelli, A. *Python in a Nutshell (In a Nutshell (O'Reilly))*; O'Reilly Media, Inc. Sebastapol, CA, 2006.

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

CT9005079

JCTC Journal of Chemical Theory and Computation

Ionization-Induced Structural Changes in Uracil Dimers and Their Spectroscopic Signatures

Anna A. Zadorozhnaya and Anna I. Krylov*

Department of Chemistry, University of Southern California, Los Angeles, California 90089-0482

Received September 29, 2009

Abstract: The electronic structure of the three representative isomers of the ionized uracil dimers is characterized by high-level electronic structure calculations. Noncovalent interactions between the fragments lower the vertical ionization energies by 0.13−0.35 eV, the largest drop being observed for the stacked and the T-shaped isomers. The initial hole is delocalized in the stacked and the H-bonded isomers and is localized in the T-shaped one. The ionization induces significant structural relaxation and increases the binding energies. The stacked dimer cation relaxes to the symmetric structure bound by 22.7 kcal/mol. The T-shaped dimer cation has a binding energy of 25.1 kcal/mol. Thus, the relative order of the stacked and T-shaped isomers is reversed upon ionization. Finally, the H-bonded isomer, which relaxes to the proton-transferred structure, is bound by 37.0 kcal/mol. The electronic spectra of all three isomers characterized at the vertical and the relaxed geometries show different patterns, which may be exploited in spectroscopic probing of ionization-induced dynamics in these species.

1. Introduction

The ionization-induced changes in DNA, which are responsible for oxidative and radiative damage of the genetic material, involve complicated coupled electron−nuclear dynamics.[1–6] The hole migration is facilitated by thermal fluctuations, which affect the ionization energies (IEs) of the individual bases, and is coupled to proton transfer. Understanding how the local environment modulates the electronic properties of nucleobases is the first step toward developing a mechanistic picture of these processes.

Gas-phase studies of small nucleobase clusters reveal the intrinsic properties of these species and allow one to quantify different effects present in realistic environments.[7] Nucleobase dimers are convenient model systems on which the effects of different types of noncovalent interactions (i.e., π-stacking, H-bonding, and electrostatics) on the electronic structure and ionization-induced dynamics can be studied by combination of state-of-the art experimental techniques and high-level theoretical methods. While the IEs of the nucleic acid bases in the gas phase have been characterized both experimentally[8–14] and computationally,[15–18] much less is known quantitatively about the effects of different interactions on the IEs in realistic environments.

Our recent combined theoretical and experimental study[19] of the homo- and heterodimers of adenine and thymine demonstrated that noncovalent interactions lower the vertical IEs by as much as 0.4 eV and that the effect is larger for thymine than for adenine. Thus, these interactions reduce the differences between the IEs of the purines and pyrimidines and promote hole migration. The magnitude and origin of the effect are different for different isomers. The largest drop in IEs was observed in the symmetric stacked and nonsymmetric H-bonded dimers. In the former case, the IE is lowered due to the hole delocalization over the two fragments and the change depends on the overlap between the fragments' molecular orbitals (MOs). In the latter case, the overlap does not play an important role: the hole, which is localized on one of the fragments, is stabilized by the electrostatic interactions with the "neutral" fragment. The magnitude of the IE drop is determined by the magnitude and relative orientation of the dipole moment of the spectator fragment. The changes of the IEs due to H-bonding in the symmetric

* Corresponding author phone: (213) 740-4929; e-mail: krylov@usc.edu.

10.1021/ct900515a
Published on Web 01/27/2010

H-bonded dimers were found to be smaller.[19] A similar trend was observed for H-bonded cytosine dimers.[20]

Similar effects of π-stacking and H-bonding on the vertical IEs of the uracil dimer have been characterized in our previous work using high-level electronic structure calculations.[21] Earlier studies of the effects of π-stacking on the IEs of nucleobases include Hartree−Fock and DFT estimates using the Koopmans theorem[22–26] and MP2 (Møller−Plesset perturbation theory) and CASPT2 (perturbatively corrected complete active space self-consistent field) calculations.[15,17,27]

In agreement with simple MO considerations, ionization changes the bonding in the dimers, resulting in significant structural relaxation. Ionization increases the binding energy, yields tighter structures, and changes the relative stability of different isomers. Moreover, in H-bonded dimers it may initiate barrierless (or almost barrierless) proton transfer, which is believed to be coupled to hole hopping.[4–6,19,22,28–30] While the ionization-induced dynamics may be very complex and its modeling requires full-dimensional coupled nuclear and electronic dynamics calculations (e.g., as in the recent study of uracil),[18] the key features of these processes can be learned from analyzing differences in the electronic states and structural parameters at the initial and the relaxed geometries (i.e., equilibrium structures of the neutral and the cation, respectively). The focus of this work is on the ionization-induced changes in the structures, binding energies, and electronic states of representative isomers of the uracil dimer. We also discuss spectroscopic signatures of the relaxation, which may be exploited in time-resolved experiments.

An interesting feature of the noncovalent dimers is the appearance of strong so-called charge-resonance (CR) bands[31] in their electronic spectra upon ionization.[32–34] These bands correspond to the transitions between the dimer molecular orbitals (DMOs) that are in-phase and out-of-phase combinations of the fragment molecular orbitals (FMOs) and are unique to the ionized dimers. Thus, they can be used as a spectroscopic probe of the ionized dimer formation. Moreover, their energies and intensities depend strongly on the overlap of FMOs and are, therefore, very sensitive to the relative orientation of and the distance between the fragments. Thus, these bands can be exploited for obtaining structural information, including ionization-induced dynamics. Other electronic transitions, which correlate with the transitions in the monomers and are called local excitations (LEs), can provide additional information. Recently, we characterized the electronic spectra in several benzene dimer isomers,[35,36] water dimers,[37,38] and two isomers of the uracil dimer.[21] While the CR bands are most intense in the symmetric dimers with favorable orbital overlap (e.g., the sandwich benzene dimer or stacked uracil dimer), they also appear in the isomers with nonequivalent fragments and more localized states (e.g., some water dimers or the T-shaped benzene dimer) where they acquire partial charge-transfer character.

Nucleobase dimers form numerous isomers.[39–42] We consider three representative isomers of the uracil dimer: H-bonded, π-stacked, and T-shaped isomers. On the neutral potential energy surface (PES), the H-bonded isomer is the lowest in energy, followed by the π-stacked and the T-shaped isomers. As demonstrated below, ionization changes the binding energies and relative ordering of the isomers. On the cation PES, the lowest energy structure corresponds to the proton-transferred H-bonded dimer, followed by the T-shaped and stacked dimers. The symmetric H-bonded cation does not have a stable minimum and undergoes barrierless proton transfer. We analyze the differences between the isomers as well as their spectroscopic signatures by using qualitative molecular orbital and electrostatic considerations, i.e., within the dimer molecular orbital−linear combination of fragment molecular orbitals (DMO−LCFMO) model.[35,37]

The structure of the paper is as follows. In section 2, we discuss the theoretical methods and computational details. In section 3 we discuss the electronic structure of the ionized dimers (section 3.1) and their equilibrium geometries (section 3.2), energetics (section 3.3), and electronic spectroscopy (section 3.4). Our concluding remarks are given in section 4. The benchmark results for density functional theory (DFT) with long-range corrected functionals augmented by empirical dispersion terms are presented in the Appendix.

2. Theoretical Methods and Computational Details

Electronic structure calculations of dimer cations are challenging owing to the open-shell character of these species. The wave function methods that are based on open-shell doublet references are often plagued by symmetry breaking and spin contamination of the underlying open-shell Hartree−Fock (HF) reference.[43,44] DFT calculations suffer from self-interaction error,[45,46] which results in artificial charge delocalization.

Within the wave function formalism, these systems are best described by the equation-of-motion coupled-cluster method for ionization potentials, EOM-IP-CCSD or simply IP-CCSD,[36,47–50] and by its less expensive configuration interaction approximation, IP-CISD.[51] EOM-IP-CCSD and IP-CISD describe problematic doublet wave functions as ionized states derived from a well-behaved closed-shell wave function; i.e., the target open-shell wave functions are generated by a Koopmans-like excitation operator R acting on the reference wave function:

$$\Psi^{\text{EOM-IP}}(N-1) = \hat{R}\Psi_0(N) \tag{1}$$

where $\Psi_0(N)$ is the wave function of the N-electron neutral system and R consists of 1h and 2h1p (one hole and two hole one particle, respectively) operators generating $(N-1)$-electron determinants from the N-electron reference. In the more accurate IP-CCSD method, Ψ_0 is a correlated CCSD wave function, whereas Ψ_0 in IP-CISD is just a single Slater determinant. The amplitudes of R are found by diagonalization of the similarity-transformed (IP-CCSD) or bare (IP-CISD) Hamiltonian.

In the DFT methods, self-interaction error can be mitigated by including long-range Hartree−Fock exchange.[52–54] We employed the ωB97X-D functional,[55] which also includes empirical dispersion terms.[56] The empirical dispersion terms

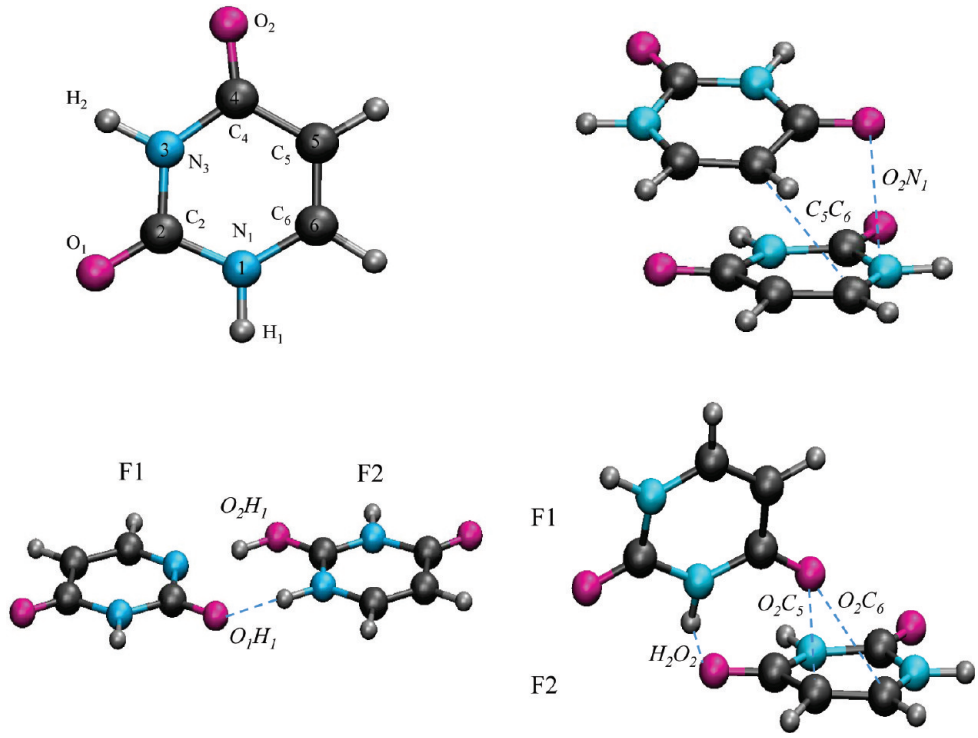


Figure 1. Definitions of the intra- and interfragment geometric parameters for uracil dimer isomers.

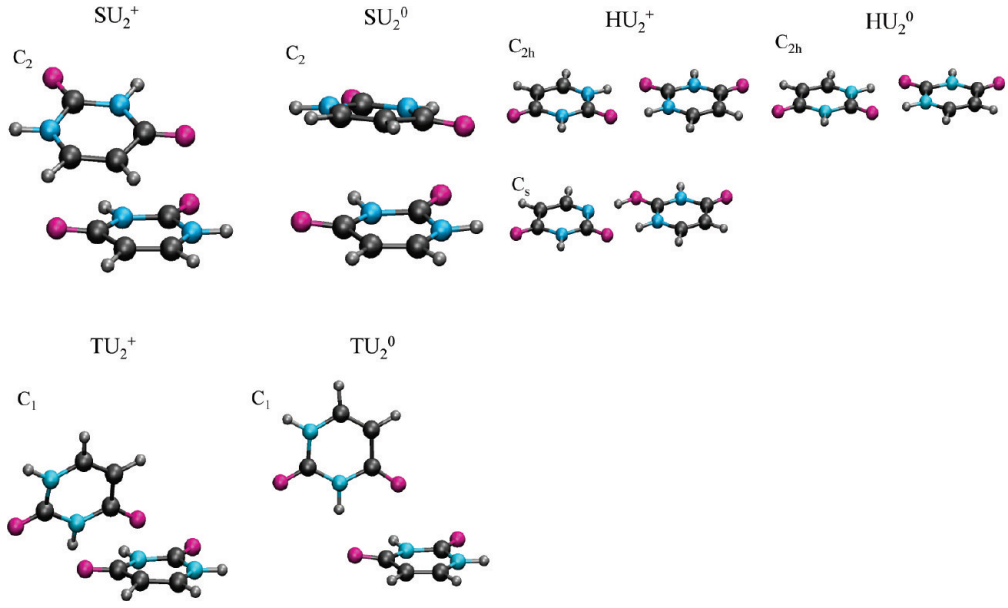


Figure 2. Geometries of the cations versus the respective neutrals for the three uracil dimer isomers.

partially mitigate the effects of basis set superposition error (BSSE) when used with an adequate basis set.

Throughout this work, we use the following notations for the isomers: HU_2, SU_2, and TU_2 refer to the H-bonded, stacked, and T-shaped isomers, respectively. For the hydrogen-bonded cations, we distinguish between the symmetric structure, which is a transition state (TS), and a proton-transferred one (PT) corresponding to the true minimum. The definitions of the inter- and intrafragment structural parameters for the stacked, T-shaped, and H-bonded isomers are given in Figure 1. The values of these parameters in the neutral and ionized systems are summarized in Tables 7 and 8. The changes in the structures induced by ionization are visualized in Figure 2.

We used EOM-IP-CCSD in calculations of the IEs, electronic spectra, and dissociation energies of the dimers, whereas for geometry optimizations and frequencies we employed IP-CISD and ωB97X-D. IP-CISD with the 6-31(+)G* basis[57] was used to optimize the SU_2^+ and HU_2^+ (TS) structures. The TU_2^+ and HU_2^+ (PT) structures were optimized with ωB97X-D and the 6-311(+)G** basis set.[58]

For both the IP-CISD and DFT-D optimizations, tight convergence criteria were enforced: the gradient and energy tolerance were set to 3×10^{-5} and 1.2×10^{-4}, respectively,

and the maximum energy change was set to 1×10^{-7}. To ensure the accuracy of the DFT-D optimizations, we employed the extrafine EML(99,590) grid.

We use the best available geometries for calculations of energy differences. The choice of the geometries is described below. In calculations of vertical properties (i.e., at the equilibrium geometries of the neutral dimers) we used the geometries from the S22 set of Hobza and co-workers.[59] The geometry of the T-shaped isomer was optimized with DFT-D as described above. To assess the possible effect of the BSSE on the structures, our study of adenine and thymine dimers[19] compared the B3LYP-D/6-31+G(d,p)-optimized structure of the stacked AT dimer versus that from the S22 set.[59] We found that the interfragment distance differs from the BSSE-corrected RI-MP2/TZVPP value[59] by only 0.076 Å. The increase of the basis set from 6-31G(d,p) to 6-311++G(2df,2pd) results in a 0.004 Å increase in interfragment separation. Thus, we do not expect significant BSSE effects on our optimized structures.

In the monomer calculations, we used the structures of the uracil cation and the neutral optimized by IP-CISD/6-31(+)G* and RI-MP2/cc-pVTZ, respectively, with the standard convergence thresholds (the gradient and energy tolerance were 3×10^{-4} and 1.2×10^{-3}, and maximum energy change was 1×10^{-6}). In all optimizations of the symmetric structures [i.e., all isomers, except for TU_2^0, TU_2^+, and HU_2^+ (PT)] the symmetry was enforced. For the stacked dimer cation we carried out an additional DFT-D optimization without the C_2 symmetry constraint that showed that the minimum indeed corresponds to the symmetric structure. In addition, vibrational analysis was performed.

For accurate energy estimates, single-point calculations were carried out at the geometries obtained as described above. The IP-CCSD method with the 6-311(+)G** basis was employed. For benchmark purposes, we also present ωB97X-D/6-311(+)G**/EML(99,590) estimates calculated at the respective DFT-D minima. The performance of different methods is discussed in the Appendix.

While the BSSE corrections can be substantial for weakly bound systems when compact basis sets are employed,[27,59,60] using augmented triple-ζ bases reduces the BSSE considerably. Moreover, empirical dispersion correction in DFT-D methods mitigates the BSSE. For example, the counterpoise correction for the binding energy in the stacked adenine−thymine dimer at the B3LYP-D/6-311+G(2df) level is only 1.4 kcal/mol.[19,61]

For the neutral stacked uracil dimer, the ωB97X-D and CCSD values of D_e are 10.5 and 11.1 kcal/mol (with the 6-311(+)G(d,p) basis set), in good agreement with the CCSD(T)/CBS value of 9.7 kcal/mol.[62] Thus, the BSSE effects are relatively small at the ωB97X-D/6-311(+)G(d,p) level even for the most problematic neutral stacked dimers. In the ionized systems, which are much more strongly bound, the effect of BSSE on the binding energy is even smaller. To quantify this effect, we computed the counterpoise correction for the stacked uracil dimer cation. The computed BSSE is 1.3 kcal/mol as estimated at the ωB97X-D level with the 6-311(+)G(d,p) basis set.

To obtain the standard thermodynamic quantities and the ZPE corrections, we performed vibrational analysis at the ωB97X-D/6-311(+)G**/EML(99,590) level for all complexes at the respective reoptimized geometries.

The electronic spectra of the dimer cations were obtained with IP-CCSD/6-31(+)G* at the cation and neutral geometries described above.

All open-shell DFT-D calculations employed the spin-unrestricted references. In these calculations, the spin contamination of the doublet Kohn−Sham determinant was low with typical $\langle S^2 \rangle$ values of 0.76−0.78.

All electrons were correlated in all the optimizations; in the single-point energy and spectral calculations the core electrons were frozen unless otherwise stated. The optimized geometries, corresponding reference energies, and frequencies are provided in the Supporting Information.

3. Results and Discussion

3.1. Molecular Orbital Framework. The character of the electronic states and the bonding patterns in ionized noncovalent dimers depend strongly on the relative orientation of the fragments.[19,21,35–37] Orbital overlap and electrostatic interactions are the most important factors determining the degree of hole delocalization, changes in bond strength due to ionization, and subsequent nuclear dynamics. When the two fragments are equivalent by symmetry, as in sandwich benzene dimers[35] or stacked C_2 nuclear base dimers,[19,21] the dimer states are derived from in-phase (bonding) and out-of-phase (antibonding) combination of the fragments MOs, and the initial hole is equally delocalized between the two fragments. The changes in IE due to dimerization depend on the orbital overlap; e.g., larger changes are observed for the states derived from ionizations of π orbitals.[19,21,35] Ionizations from antibonding orbitals increase the formal interfragment bond order and produce more tightly bound structures, whereas ionizations from the bonding orbitals result in dissociative states.

The orbital picture, changes in the vertical IEs, and initial hole delocalization are similar in symmetric hydrogen-bonded dimers; however, the ionization-induced dynamics is more complex and involves proton transfer.[19,20] The changes in the vertical IEs are smaller for most of the states due to a less favorable overlap. In dimers with nonequivalent fragments, the MOs (and, consequently, the initial hole) become more localized; however, changes in the IEs and wave functions can also be explained by overlap considerations within the DMO−LCFMO framework.[36,37] Finally, in nonsymmetric H-bonded dimers electrostatic interactions become more important than orbital overlap. For example, we observed large changes (0.4−0.7 eV) in the IEs and binding energies in some nonsymmetric hydrogen-bonded dimers of thymine and cytosine.[19,20] In these dimers, the hole localized on one of the fragments is stabilized by the dipole moment of the neutral fragment.

The electronic structure of the stacked and symmetric H-bonded uracil dimers at the respective neutral geometries was discussed in detail in ref 21. Below we focus on the T-shaped isomer. The principal difference between the

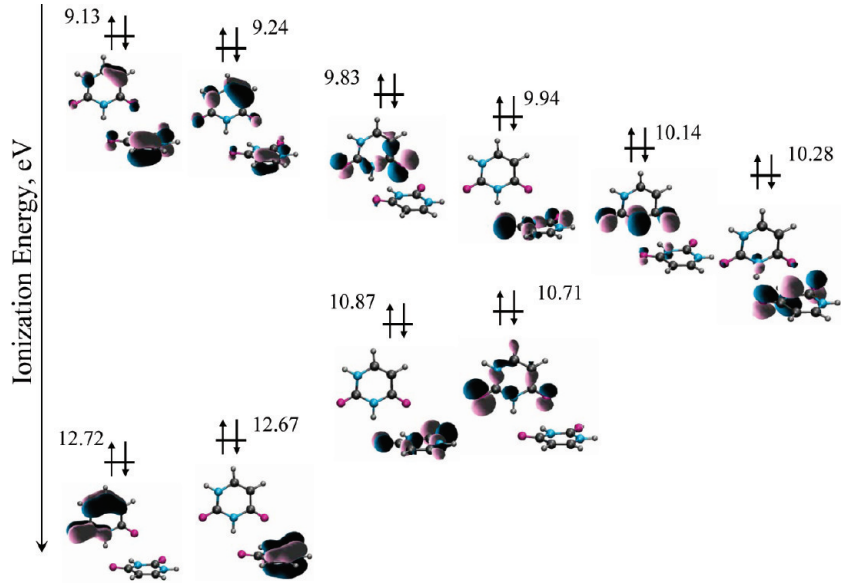


Figure 3. Ten lowest ionized states of the T-shaped uracil dimer at the neutral geometry calculated with IP-CCSD/6-311(+)G**.

T-shaped and the stacked or H-bonded structures is that in the former the two fragments are not equivalent by symmetry, which affects the electronic structure. The 10 lowest ionized states of the T-shaped uracil dimer and the corresponding MOs are presented in Figure 3. As in the stacked and H-bonded systems, the dimer MOs are formed from the MOs of the fragments, and the ionized states of the dimer correlate well with the states of the monomer (i.e., no mixing of the MOs of different character is observed). For example, the two highest lying MOs are the linear combinations of the π_{CC} MOs of the fragments. However, the MOs of the T-shaped dimer are more localized. For example, the lp(O) MO of the dimer is a localized lp(O) orbital of one of the fragments. For the four delocalized dimer orbitals [formed by the π_{CC} and lp(O) + lp(N) fragment orbitals] the distribution of electron density is also uneven. Owing to a less favorable overlap between the fragment MOs, the splitting between the pairs of ionized states in the T-shaped dimer is smaller. The largest splitting of 0.14 eV was observed for the dimer states derived from the π-like lp(O) + lp(N) fragment orbitals.

Despite less efficient overlap and smaller splittings between the pairs of states derived from the same FMOs, the absolute changes in the IEs in the T-shaped isomer are similar to those in the stacked dimer. For example, the lowest IE of this isomer is 9.13 eV. This value is red-shifted by 0.35, 0.22, and 0.01 eV relative to the first IE of the monomer and symmetric H-bonded and π-stacked dimers, respectively. This is similar to large changes in the IEs observed in the nonsymmetric H-bonded dimers of thymine and cytosine, where lowering of the IE was due to electrostatic stabilization of the localized hole by the dipole moment of the "neutral" fragment. The dipole moment of uracil is 4.19 D, which is comparable to the dipole moment of thymine (4.11 D).

3.2. Ionization-Induced Structural Changes: Equilibrium Geometries of the Uracil Dimer Cations. Ionization induces significant structural changes in the dimers, as can be seen from Figure 2. In the analysis below, we distinguish between the changes in the structures of the fragments (and compare those to ionization-induced changes in the monomer) and the interfragment relaxation. The definitions of the parameters are given in Figure 1, and their values are

Table 1. Values of the Optimized Structural Parameters (Å, deg) of the Fragments in the Stacked, H-Bonded, H-Transferred H-Bonded, and T-Shaped Uracil Dimer Cations[a]

param	SU_2^+	HU_2^+ (TS)	HU_2^+ (PT), F1	HU_2^+ (PT), F2	TU_2^+, F1	TU_2^+, F2	U^+
C_4-C_5	1.461, +0.010	1.461, +0.011	1.461, +0.011	1.458, +0.008	1.431, −0.026	1.475, +0.024	1.457, +0.011
C_5-C_6	1.367, +0.018	1.352, +0.002	1.407, +0.057	1.337, −0.013	1.353, +0.011	1.392, +0.050	1.386, +0.043
C_6-N_1	1.330, −0.038	1.352, −0.017	1.310, −0.059	1.391, +0.022	1.357, −0.012	1.324, −0.045	1.316, −0.049
N_1-C_2	1.405, +0.023	1.379, +0.012	1.411, +0.044	1.332, −0.035	1.389, −0.002	1.429, +0.044	1.433, +0.053
C_2-N_3	1.368, −0.014	1.349, −0.022	1.363, −0.008	1.331, −0.040	1.401, +0.023	1.377, −0.003	1.357, −0.017
N_3-C_4	1.384, −0.017	1.399, −0.008	1.400, −0.007	1.438, +0.031	1.365, −0.032	1.384, −0.007	1.387, −0.010
C_4-O_2	1.198, −0.024	1.190, −0.028	1.204, −0.014	1.194, −0.024	1.257, +0.041	1.206, −0.014	1.195, −0.020
C_2-O_1	1.182, −0.034	1.208, −0.023	1.216, −0.015	1.287, +0.056	1.195, −0.012	1.190, −0.017	1.178, −0.034
$C_4-C_5-C_6$	119.3, −0.5	119.5, −0.1	119.4, −0.2	120.4, +0.7	118.4, −1.1	119.5, +0.3	119.7, −0.1
$C_5-C_6-N_1$	121.0, −0.9	121.1, −1.5	123.1, +0.6	121.8, −0.7	121.9, +0.2	120.1, −1.8	119.4, −2.6
$C_6-N_1-C_2$	124.3, +0.8	123.4, +0.9	120.1, −2.4	121.0, −1.5	123.7, +0.2	124.9, +1.4	125.5, +2.0
$N_1-C_2-N_3$	113.8, +0.8	115.4, +1.1	118.2, +3.9	118.8, +4.5	112.9, −0.6	113.5, +0.4	113.6, +0.8
$C_2-N_3-C_4$	126.9, −1.2	126.3, −1.8	125.5, −2.6	125.5, −2.6	126.2, −1.1	127.0, −0.4	126.2, −2.4
$N_3-C_4-C_5$	114.7, +1.3	114.3, +1.4	113.7, +0.8	112.5, −0.4	116.9, +2.5	114.7, +0.2	115.7, +2.4
Σ(angles)	719.9, +0.3				720.0, +0.0	719.7, +0.1	720.0, +0.0

[a] The differences (Å, deg) with respect to the equilibrium geometry of the respective neutral complex are also given, showing the ionization-induced changes in the geometry. See Figure 1 for the definitions of the parameters.

Table 2. Values of the Interfragment Structural Parameters (Å, deg) of the Stacked, H-Bonded, H-Transferred H-Bonded, and T-Shaped Uracil Dimer Cations[a]

SU_2^+		HU_2^+ (TS)		HU_2^+ (PT)		TU_2^+	
C_5-C_6	3.299 (−0.451)	O_1-H_1	1.828 (+0.053)	O_1-H_1	1.749 (−0.026)	H_2-O_2	2.000 (+0.072)
O_2-N_1	3.116 (−0.175)	O_2-H_1	1.828 (+0.053)	O_2-H_1	1.018 (−0.757)	O_2-C_5	2.178 (−1.099)
						O_2-C_6	2.701 (−0.950)
α	18.4 (+5.6)						
d	3.51 (+0.34)						

[a] The differences (Å, deg) with respect to the equilibrium geometry of the respective neutral complexes are given in parentheses. See Figure 1 for the definitions of the parameters.

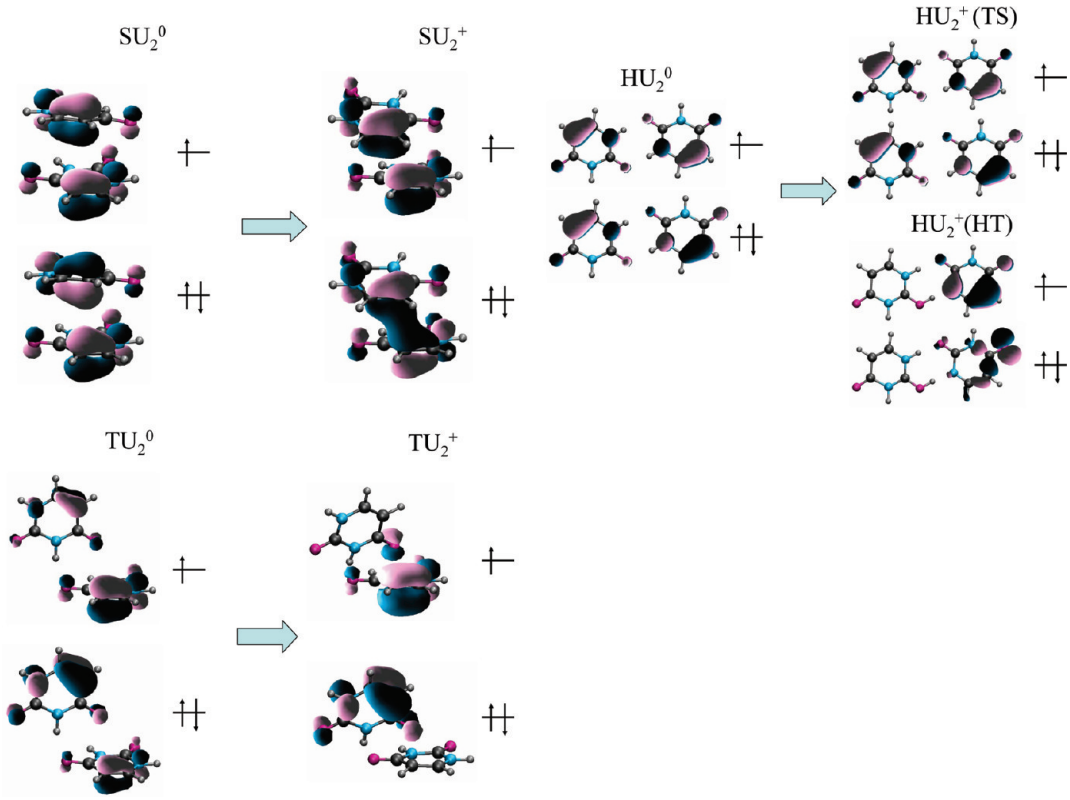


Figure 4. Two highest occupied MOs of the three isomers of the uracil dimer at the neutral and cation geometries.

summarized in Tables 1 and 2. Only the symmetry-unique parameters are given.

First, let us consider the effect of ionization on the intrafragment parameters (see Table 1) and compare the monomer and the symmetric dimer cation data. The magnitude of relaxation in the monomer is larger than in the stacked and H-bonded dimers. For instance, the C_5-C_6 bond increases by 0.043 Å in the monomer versus 0.018 and 0.002 Å in the stacked and H-bonded dimers, respectively. The sign of the change in the monomer and the symmetric dimers is the same for all the parameters, which is consistent with the DMO−LCFMO picture. The magnitude of the changes is smaller in the dimers because the hole is delocalized over the two fragments.

In the nonsymmetric dimers, the fragments are not equivalent and the orbital picture is more complicated. The hole is distributed unevenly between the two fragments, such that the positive charge is localized on one of them. Comparing the data presented in Table 7 for the H-bonded proton-transferred and the T-shaped dimer cations with those of the monomer, we observe that the structural changes of fragment 1 (F1) of HU_2^+ (PT), fragment 2 (F2) of TU_2^+, and the monomer cation are very similar. For instance, the C_5-C_6 bond increases by 0.057, 0.050, and 0.043 Å in fragment 1 of HU_2^+ (PT), fragment 2 of TU_2^+, and the monomer cation, respectively. Thus, one of the fragments in nonsymmetric dimers relaxes similarly to the monomer cation, while the other adjusts accordingly. This is similar to what is found in the T-shaped benzene dimer.[36]

The ionization-induced changes in the interfragment parameters (given in Table 2) and the MOs (shown in Figure 4) are consistent with the DMO−LCFMO predictions: the fragments adjust their relative orientation to maximize the overlap between their HOMOs (π_{CC}).

The change in the MOs is illustrated in Figure 4 depicting HOMOs at the neutral and the cation geometries. In the stacked dimer, the two π_{CC} FMOs give rise to efficient overlap, lending a partial covalent character to the ionized dimer. In the T-shaped dimer, the changes in the HOMO are different. Upon relaxation, the hole becomes more localized on the lower fragment, and the only contribution to the overlap is due to the oxygen lone pair of the top fragment pointing toward the π_{CC} MO of the lower fragment.

Table 3. Total (E_{tot}, hartrees) and Dissociation (D_e, kcal/mol) Energies of the Four Isomers of the Uracil Dimer in the Neutral and Ionized States Computed by CCSD/IP-CCSD with 6-311(+)G** [a]

complex	E_{tot}^{CCSD}	D_e^{CCSD}	ΔE^{CCSD}
U^0	−413.882 346		
U^+	−413.542 383		−5.41
UH^+	−414.209 422		
$(U-H)^0$	−413.212 558		
SU_2^0	−827.782 419	11.1	
SU_2^+	−827.456 874	20.2	−6.48
HU_2^0	−827.793 226	17.9	
HU_2^+ (TS)[b]	−827.450 565	16.2	−0.64
HU_2^+ (PT)[c]	−827.475 648	32.0/33.7	
TU_2^0	−827.779 232	9.1	
TU_2^+	−827.463 991	24.6	−12.71

[a] The relevant total energies of the uracil monomer are also given. The relaxation energies (ΔE, kcal/mol) defined as the difference in the total energies of the cation at the neutral and relaxed cation geometries are also shown. For HU_2^+ (PT) dissociation energies corresponding to the $U^0 + U^+/(U-H)^0 + UH^+$ channels are given. [b] Transition state. [c] Proton-transferred structure, UH^+ (U−H).

The magnitude of the relaxation is quantified by Table 3, which presents the differences in the total energies between the relaxed and vertical structures of the dimer cations calculated by EOM-IP-CCSD/6-311(+)G**. For the T-shaped, stacked, and H-bonded isomers, ΔE^{CCSD} is −12.71, −6.48, and −0.64 kcal/mol, respectively. Such a large relaxation effect in the T-shaped cation is somewhat surprising, as from Figure 4 the FMOs overlap more efficiently in the stacked dimer. The reason is the electrostatic interaction of the lone pair on the oxygen of fragment 1 and the hole on fragment 2, which stabilizes the T-shaped structure.[19]

The interfragment parameters presented in Table 2 are consistent with the MO changes. In the stacked dimer cation, the fragments slide with respect to each other, so the overlap of FMOs centered on the C_5, C_6, N_1, and O_2 atoms increases (see Figure 4). The C_5-C_6 and O_2-N_1 distances decrease by 0.451 and 0.175 Å, respectively. Surprisingly, the distance between the centers-of-masses of the fragments increases by 0.34 Å in the cation with respect to the neutral. This illustrates that the average geometric parameters in polyatomic systems can be misleading.

In the T-shaped cation, the fragments move to minimize the distance between the lone pair on O_2 of the top fragment and the π_{CC} MO of the bottom fragment. The characteristic parameters in this case are the O_2-C_5 and O_2-C_6 distances, which decrease by 1.099 and 0.950 Å, respectively.

In the symmetric H-bonded dimer, the structural changes and, consequently, relaxation energy are small. As one can see from Figure 4, there are also no significant changes in the MOs upon relaxation due to unfavorable orbital overlap. Moreover, this symmetric structure is a transition state, as shown by the vibrational analysis discussed later. Much larger stabilization is achieved by a proton transfer, which lowers the total energy by 15.7 kcal/mol, making the proton-transferred H-bonded isomer the lowest energy structure on the cation's PES.

3.3. Binding Energies of the Neutral and Ionized Uracil Dimers: Potential and Free Energy Calculations. *3.3.1. Potential Energy Profile.* Figures 5 and 6 present the relative ordering and binding energies of the neutral and ionized uracil dimers calculated by IP-CCSD and ωB97X-D with the 6-311(+)G** basis. In the neutral, the symmetric H-bonded uracil dimer is the minimum energy isomer, with the stacked and T-shaped dimers lying 6.8 and 8.8 kcal/mol higher in energy. Excluding the proton-transferred dimer, the lowest energy cation structure is the T-shaped one. The energy spacing between the T-shaped and the stacked and H-bonded cations is 4.4 and 8.4 kcal/mol, respectively. Upon proton transfer the total energy of the H-bonded cation is lowered by 15.8 kcal/mol, so that it lies 7.4 kcal/mol below that of the T-shaped cation.

The calculated binding energies for the H-bonded, stacked, and T-shaped neutral dimers are 17.9, 11.1, and 9.1 kcal/mol, respectively. The DFT-D and CCSD values are within 1 kcal/mol of each other. The D_e values for the stacked and

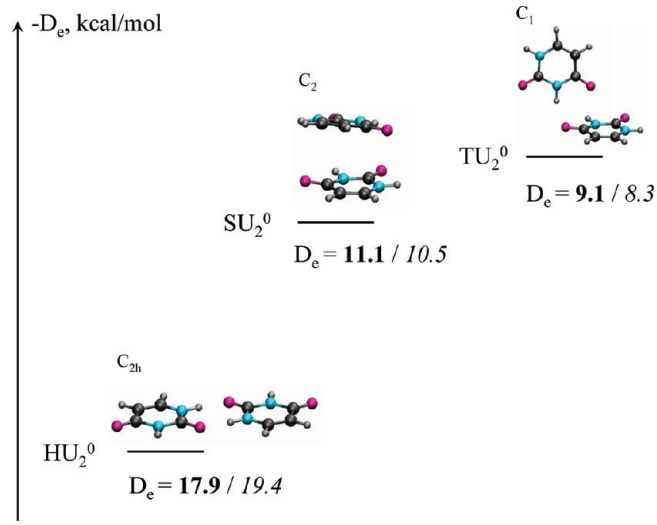


Figure 5. Binding energies (kcal/mol) of the three isomers of the neutral uracil dimer calculated at two levels of theory: CCSD/6-311(+)G** (shown in bold) and ωB97X-D/6-311(+)G**/EML(99,590) (shown in italic).

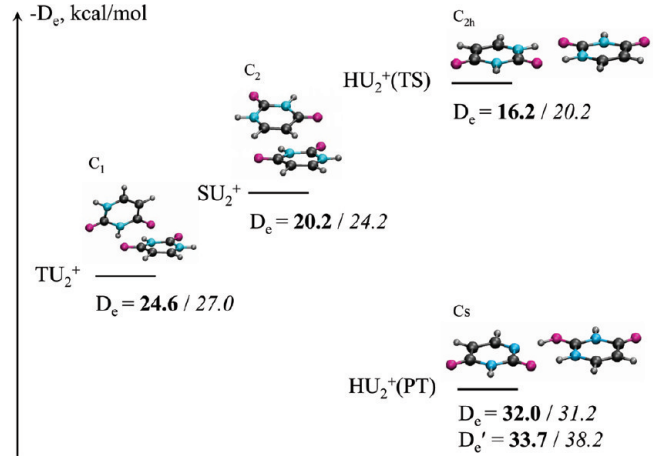


Figure 6. Binding energies (kcal/mol) of the three isomers of the uracil dimer cation calculated at two levels of theory: IP-CCSD/6-311(+)G** (shown in bold) and ωB97X-D/6-311(+)G**/EML(99,590) (shown in italic). For the proton-transferred H-bonded uracil dimer cation, the binding energies corresponding to the two dissociation limits are presented.

Table 4. Dissociation Energies (kcal/mol) and Standard Thermodynamic Quantities of the Neutral and the Cation Uracil Dimers Calculated at the ωB97X-D/6-311(+)G**/EML(99,590) Level[a]

reaction	D_e	D_0	$\Delta H°$, kcal/mol	$\Delta S°$, cal/(mol K)	$\Delta G°$, kcal/mol
$SU_2^0 \rightarrow U^0 + U^0$	10.5	9.8	8.4	31.5	−1.0
$SU_2^+ \rightarrow U^0 + U^+$	24.4	22.7	20.9	40.4	8.8
$HU_2^0 \rightarrow U^0 + U^0$	19.4	18.2	16.8	38.1	5.4
HU_2^+ (TS) $\rightarrow U^0 + U^+$	20.2	21.8	23.2	40.5	11.1
HU_2^+ (TS) $\rightarrow HU_2^+$ (PT)	11.0	13.1	−8.8	2.7	−9.6
HU_2^+ (PT) $\rightarrow U^0 + U^+$	31.2	30.6	−0.7	37.7	18.7
HU_2^+ (PT) $\rightarrow (U - H)^0 + UH^+$	38.2	37.0	−1.3	38.6	24.2
$TU_2^0 \rightarrow U^0 + U^0$	8.3	7.6	6.2	29.6	−2.6
$TU_2^+ \rightarrow U^0 + U^+$	27.0	25.1	23.0	38.8	11.4

[a] For the proton-transferred H-bonded cation the values corresponding to the two different dissociation limits are given.

H-bonded isomers are also in good agreement with the recent CCSD(T)/CBS values of 20.4 and 9.7 kcal/mol from ref 62.

Note that the interaction of the fragments in the neutral uracil dimers is much stronger than in the benzene dimers, where the typical interaction energies lie in the range of 1.5−3.0 kcal/mol for all isomers.[63,64] The binding energies increase upon ionization, in agreement with the DMO−LCFMO predictions. In the T-shaped, stacked, and symmetric H-bonded cations the fragments are bound by 24.6, 20.2, and 16.2 kcal/mol, respectively. For comparison, in the benzene dimer cation the binding energies are 20 and 12 kcal/mol for the sandwich and T-shaped isomers, respectively.[35,36] However, the strongest interaction is observed in the proton-transferred H-bonded cation, where the binding energy corresponding to the $U^0 + U^+$ dissociation channel is 32.0 kcal/mol [this channel lies 1.8 kcal/mol below an alternative $(U-H)^0 + UH^+$ channel].

In conclusion, when the uracil dimer is ionized, the interaction between the fragments increases almost 2-fold for the stacked and H-bonded isomers and more than 2-fold for the T-shaped isomer. Such a strong increase in interaction in the T-shaped structure is very different from that of the benzene dimer cation and can be explained by electrostatic interactions rather than orbital overlap considerations. The H-bonded isomer is stabilized by the proton transfer.

3.3.2. Free Energy Profile. It has been argued that the entropy contribution to the stability can be important in the nucleobase dimer systems favoring stacked isomers over H-bonded isomers.[65] Thus, we performed vibrational analysis using ωB97X-D. Moreover, we wanted to quantify the zero-point energy (ZPE) corrections to the dissociation energies. The calculated dissociation energies and the standard thermodynamic quantities for the dissociation of the neutral and the ionized dimers are given in Table 4.

Among the neutral uracil dimers, only the H-bonded isomer is predicted to be stable under the standard conditions ($\Delta G° = 5.4$ kcal/mol). Standard Gibbs free energies, $\Delta G°$, of the stacked and T-shaped isomers are −1.0 and −2.6 kcal/mol, respectively. The data in Table 2 show that the entropy contribution is similar for all three isomers: $\Delta S°$ of dissociation is 31.5, 38.1, and 29.6 cal/(mol K) for the stacked, H-bonded, and T-shaped isomers, respectively. However, more appropriate treatment including anharmonicities may discriminate between the isomers more. The enthalpy contribution is different: for the H-bonded uracil dimer the enthalpy of dissociation is 16.8 kcal/mol, whereas the corresponding values for the stacked and T-shaped isomers are 8.4 and 6.2 kcal/mol, respectively.

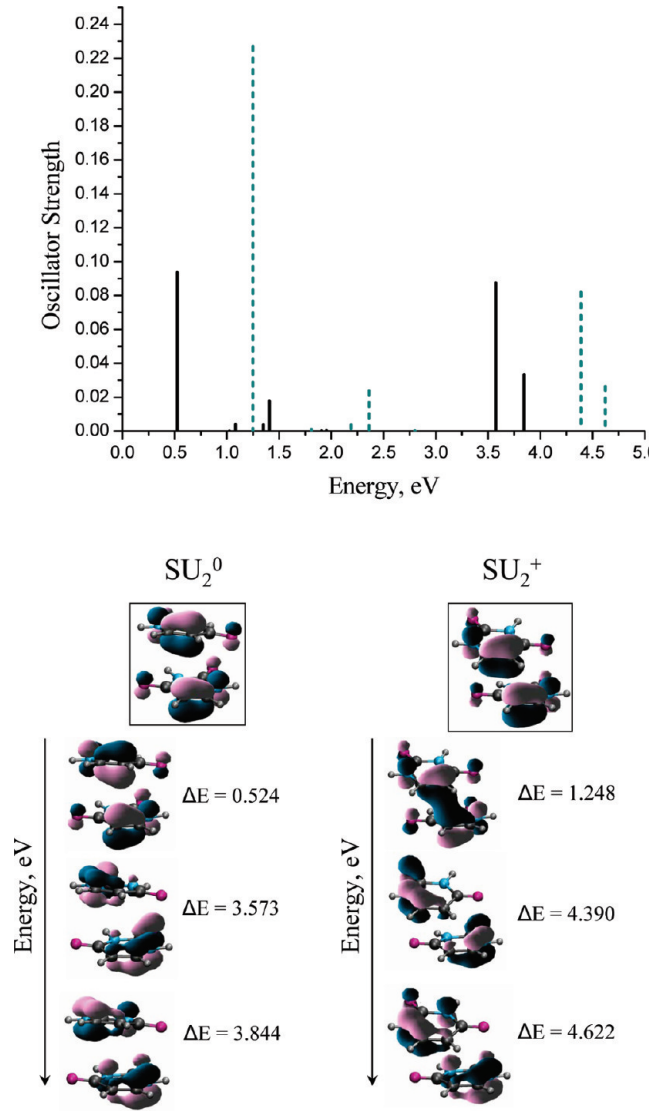


Figure 7. Electronic spectra (top panel) of the stacked uracil dimer cation at the neutral (solid black) and the cation (dashed blue) geometries calculated with IP-CCSD/6-31(+)G* and the electronic states corresponding to the three most intense transitions (bottom panel).

Unlike neutrals, all of the dimer cation isomers are stable under the standard conditions. The most stable isomer is the proton-transferred H-bonded cation with a $\Delta G°$ of 18.7 kcal/mol. In order of decreasing stability, the proton-transferred

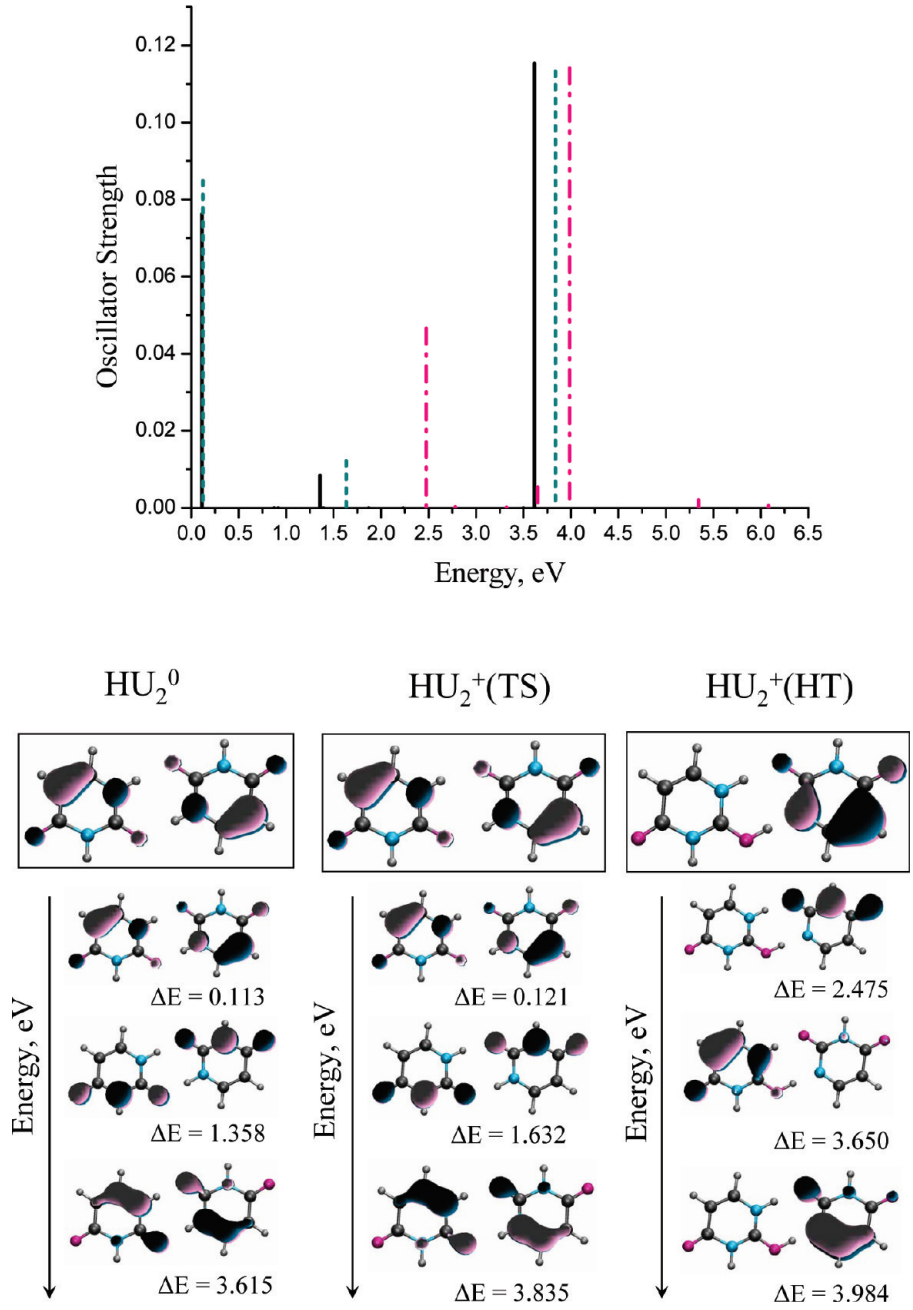


Figure 8. Electronic spectra (top panel) of the H-bonded uracil dimer cation at the neutral (solid black), symmetric transition state (dashed blue), and proton-transferred cation (dashed−dotted pink) geometries calculated with IP-CCSD/6-31(+)G* and the electronic states corresponding to the three most intense transitions (bottom panel).

dimer is followed by the T-shaped, symmetric H-bonded (TS), and stacked isomers. Again, the $\Delta S°$ values are very close for all of the isomers, being 40.4, 40.5, 37.7, and 38.8 cal/(mol K) for SU_2^+, HU_2^+ (TS), HU_2^+ (PT), and TU_2^+, respectively, whereas the $\Delta H°$ contributions are different.

Thus, we conclude that the enthalpy determines the relative stability of the neutral and ionized uracil dimers to a high degree, while the entropy contribution has a less pronounced effect.

Lastly, the ZPE corrections lower the dissociation energy estimates by 0.6−1.9 kcal/mol for all the neutral and ionized dimers, except for the symmetric H-bonded dimer. In the symmetric H-bonded dimer, the ZPE correction has the opposite sign and increases the dissociation energy by 1.6 kcal/mol, because this structure is a transition state with one imaginary frequency.

3.4. Electronic Spectra of the Uracil Dimer Cations. This section presents the calculated electronic spectra of the uracil dimer cations. The spectra of the stacked and H-bonded isomers at the geometry of the neutral were described in a detail in previous work;[21] therefore, we focus on the effect of geometry relaxation on the spectroscopic properties. For the H-bonded dimer, we present the spectra of both the symmetric (TS) and the proton-transferred structures.

Figures 7−9 present the electronic spectra of the stacked, H-bonded, and T-shaped uracil dimers, respectively, calculated by IP-CCSD/6-31(+)G* at the neutral and the cation geometries. Figures 7−9 also show the character of the

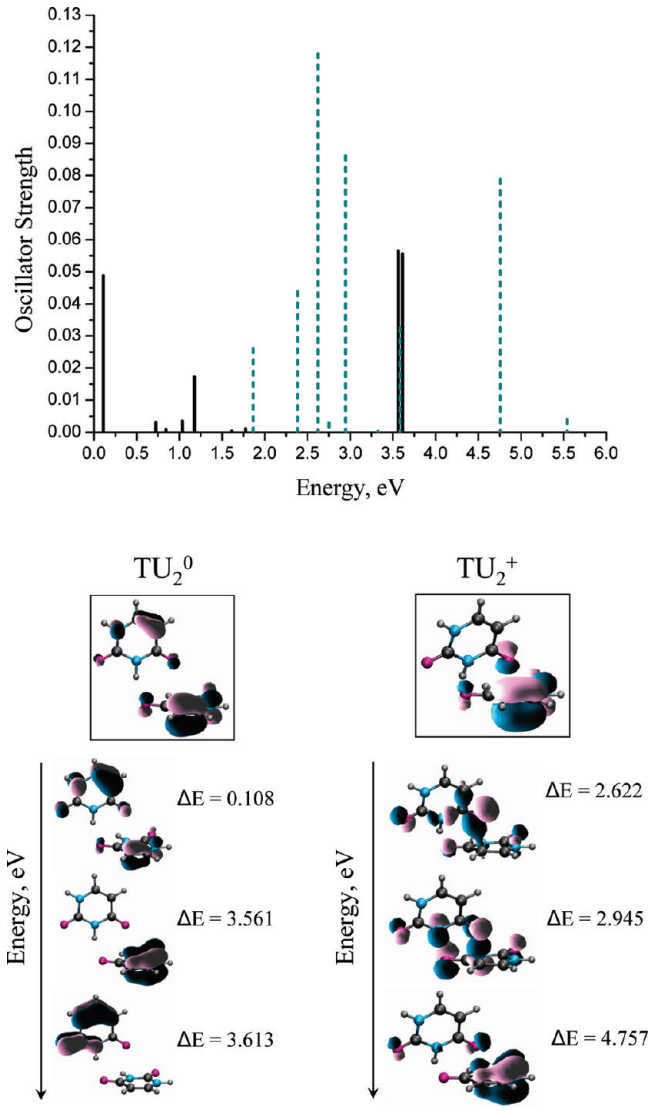


Figure 9. Electronic spectra (top panel) of the T-shaped uracil dimer cation at the neutral (solid black) and the cation (dashed blue) geometries calculated with IP-CCSD/6-31(+)G* and the electronic states corresponding to the three most intense transitions (bottom panel).

electronic states corresponding to the three most intense transitions in each spectrum. The transition energies, transition dipole moments, and oscillator strengths are provided in Tables 5−8.

The spectrum of the stacked dimer at the neutral geometry is dominated by three intense lines at 0.5, 3.5, and 3.8 eV (see Figure 7). The first peak is the CR band, which is unique to the dimer, while the others are the LEs between the states of the cation with the various π-orbitals singly occupied. Upon geometric relaxation, the spectrum shifts to higher energies by approximately 0.8 eV, so the lines appear at 1.2, 4.4, and 4.6 eV. The intensity of the charge resonance band increases more than 2-fold upon relaxation.

The H-bonded dimer cation spectrum at the geometry of the neutral (see Figure 8) features two intense lines at 0.1 and 3.6 eV and a small peak at 1.3 eV. As in the stacked cation, these lines are the CR band and two LEs corresponding to the transition between the π-orbitals of the cation (see Figure 8). The CR band is less intense than in the stacked

Table 5. Excitation Energies (ΔE, eV), Transition Dipole Moments ($\langle\mu^2\rangle$, au), and Oscillator Strengths (f) of the Stacked Dimer Cation at the Geometry of the Neutral and Cation (IP-CCSD/6-31(+)G*)

	neutral			cation		
transition	ΔE	$\langle\mu^2\rangle$	f	ΔE	$\langle\mu^2\rangle$	f
$X^2B \rightarrow 1^2A$	0.524	7.2918	0.0935	1.248	7.4212	0.2269
$X^2B \rightarrow 2^2B$	1.023	0.0028	0.0000	1.799	0.0010	0.0000
$X^2B \rightarrow 2^2A$	1.081	0.1503	0.0040	1.809	0.0197	0.0009
$X^2B \rightarrow 3^2B$	1.349	0.1141	0.0038	2.190	0.0709	0.0038
$X^2B \rightarrow 3^2A$	1.406	0.5171	0.0178	2.362	0.4090	0.0237
$X^2B \rightarrow 4^2B$	1.906	0.0024	0.0001	2.798	0.0010	0.0000
$X^2B \rightarrow 4^2A$	1.952	0.0053	0.0003	2.800	0.0016	0.0001
$X^2B \rightarrow 5^2B$	3.573	0.3531	0.0333	4.390	0.7613	0.0819
$X^2B \rightarrow 5^2A$	3.844	0.9990	0.0875	4.622	0.2323	0.0263

Table 6. Excitation Energies (ΔE, eV), Transition Dipole Moments ($\langle\mu^2\rangle$, au), and Oscillator Strengths (f) of the Symmetric H-Bonded Dimer Cation at the Geometry of the Neutral and Cation (IP-CCSD/6-31(+)G*)

	neutral			cation		
transition	ΔE	$\langle\mu^2\rangle$	f	ΔE	$\langle\mu^2\rangle$	f
$X^2A_u \rightarrow 1^2B_g$	0.113	27.4607	0.0763	0.121	28.7406	0.0849
$X^2A_u \rightarrow 1^2B_u$	0.871	0.0000	0.0000	1.064	0.0000	0.0000
$X^2A_u \rightarrow 1^2A_g$	0.915	0.0003	0.0000	1.123	0.0003	0.0000
$X^2A_u \rightarrow 2^2B_g$	1.358	0.2527	0.0084	1.632	0.3048	0.0122
$X^2A_u \rightarrow 2^2A_u$	1.391	0.0000	0.0000	1.683	0.0000	0.0000
$X^2A_u \rightarrow 2^2B_u$	1.867	0.0000	0.0000	1.954	0.0000	0.0000
$X^2A_u \rightarrow 2^2A_g$	2.232	0.0000	0.0000	2.381	0.0000	0.0000
$X^2A_u \rightarrow 3^2A_u$	3.501	0.0000	0.0000	3.740	0.0000	0.0000
$X^2A_u \rightarrow 3^2B_g$	3.615	1.3026	0.1154	3.835	1.2053	0.1133

Table 7. Excitation Energies (ΔE, eV), Transition Dipole Moments ($\langle\mu^2\rangle$, au), and Oscillator Strengths (f) of the H-Bonded Dimer Cation at the Optimized Proton-Transferred Geometry (IP-CCSD/6-31(+)G*)

transition	ΔE	$\langle\mu^2\rangle$	f
$X^2A'' \rightarrow 1^2A'$	1.702	0.0004	0.0000
$X^2A'' \rightarrow 2^2A''$	2.475	0.7690	0.0466
$X^2A'' \rightarrow 2^2A'$	2.782	0.0040	0.0003
$X^2A'' \rightarrow 3^2A'$	3.325	0.0024	0.0002
$X^2A'' \rightarrow 3^2A''$	3.650	0.0605	0.0054
$X^2A'' \rightarrow 4^2A''$	3.984	1.1704	0.1142
$X^2A'' \rightarrow 4^2A'$	4.493	0.0001	0.0000
$X^2A'' \rightarrow 5^2A''$	5.343	0.0162	0.0021
$X^2A'' \rightarrow 5^2A'$	6.082	0.0039	0.0006

cation, and the most intense transition is the LE at 3.6 eV. The spectrum at the transition-state structure exhibits only minor differences, i.e., 0.1 eV blue shifts in the peak positions with the intensities remaining the same. However, the spectrum and the character of the states change dramatically upon proton transfer. A new band appears at 2.5 eV. The localized character of the states and C_s symmetry make the proton-transferred H-bonded cation spectrum very similar to that of the uracil cation.

In the T-shaped cation spectrum at the neutral geometry, the CR and the two intense LE transitions appear at 0.1, 3.5, and 3.6 eV (see Figure 9). The spectrum is very similar to that of the H-bonded isomer at the neutral geometry. As in the stacked and H-bonded cations, the transitions between the π-like orbitals are the most intense. However, the character of the states is different: the states are more

Table 8. Excitation Energies (ΔE, eV), Transition Dipole Moments ($\langle\mu^2\rangle$, au), and Oscillator Strengths (f) of the T-Shaped Dimer Cation at the Geometry of the Neutral and Cation (IP-CCSD/6-31(+)G*)

	neutral			cation		
transition	ΔE	$\langle\mu^2\rangle$	f	ΔE	$\langle\mu^2\rangle$	f
$X^2A_1 \rightarrow 2^2A_1$	0.108	18.4996	0.0488	1.866	0.5715	0.0261
$X^2A_1 \rightarrow 3^2A_1$	0.725	0.1761	0.0031	2.384	0.7506	0.0438
$X^2A_1 \rightarrow 4^2A_1$	0.841	0.0436	0.0009	2.622	1.8376	0.1180
$X^2A_1 \rightarrow 5^2A_1$	1.031	0.1376	0.0035	2.750	0.0428	0.0029
$X^2A_1 \rightarrow 6^2A_1$	1.176	0.5961	0.0172	2.945	1.1927	0.0861
$X^2A_1 \rightarrow 7^2A_1$	1.609	0.0095	0.0004	3.324	0.0042	0.0003
$X^2A_1 \rightarrow 8^2A_1$	1.776	0.0261	0.0011	3.584	0.3711	0.0326
$X^2A_1 \rightarrow 9^2A_1$	3.561	0.6475	0.0565	4.757	0.6759	0.0788
$X^2A_1 \rightarrow 10^2A_1$	3.613	0.6276	0.0555	5.539	0.0295	0.0040

localized. Upon relaxation, the spectrum changes completely, as does the character of the states. The maximum intensity increases 2.5 times, and new intense lines appear in the 1.7−3.0 and 4.5−5.0 eV regions. The orbital picture is now much more complex: the DMOs become combinations of several FMOs. Thus, the electronic transitions can no longer be described as CR or LE excitations. The most intense bands correspond to the transitions between the cation states with the π_{CC} orbital and the lp(O) orbital singly occupied and are of charge-transfer character.

To summarize, the three isomers have distinctly different spectra, which can be used to distinguish between them experimentally. Moreover, significant changes upon relaxation may be exploited to monitor ionization-induced dynamics in a pump−probe experiment. Immediately upon ionization, the isomers will exhibit intense lines in three regions: 0.0−0.7, 1−1.5, and 3.0−4.0 eV. While the spectra of the H-bonded and T-shaped dimers at the neutral geometry are similar, the stacked cation can be distinguished by the two peaks of moderate intensity in the 0.5−0.7 and 3.5−4.0 eV regions. Upon relaxation, the most intense CR band of the stacked isomer shifts to 1.2 eV and acquires additional intensity. The relaxation of the T-shaped cation manifests itself by significant growth of intensity in the 2.5−3.0 eV region. The hydrogen-bonded complex is more difficult to distinguish because of the overlap of its spectral lines with the stacked and T-shaped spectra. Still, the signature of proton transfer is the 0.3−0.4 eV blue shift of the intense transition in the 3.5−4.0 eV region.

4. Conclusions

We characterized the electronic structure of three representative isomers of the ionized uracil dimers: H-bonded, stacked, and T-shaped. The interactions between the fragments lower the vertical IEs by 0.13−0.35 eV, the largest drop in IE being observed for the stacked and T-shaped isomers. Interestingly, the character of the ionized states and the origin of the IE change are different in these two isomers. In the stacked dimer, the hole is delocalized between the two fragments and orbital overlap determines the change in the IE. In the T-shaped isomer, the hole is localized and the change in the IE is due to electrostatic interactions between the "ionized" and the "spectator" fragments. The change in the IE for the symmetric H-bonded dimer is small, because neither overlap nor electrostatic interactions can stabilize the hole; however, larger changes are expected for the nonsymmetric H-bonded dimers.[19]

The geometric relaxation is also different for the three isomers. The stacked isomer relaxes to a tighter structure with more efficient overlap between the FMOs, and the hole remains delocalized between the fragments. The H-bonded isomer undergoes proton transfer, forming the lowest energy structure on the cation's surface in which the charge and the unpaired electron are localized on different moieties. Finally, the T-shaped dimer relaxes to the structure with the localized hole. The respective binding energies of the cation isomers are 22.7, 37.0, and 25.1 kcal/mol.

Finally, we characterized the electronic spectra of the cations at the neutral and the relaxed geometries. At the neutral geometry, the H-bonded and stacked isomers feature intense CR bands at 0.1 and 0.5 eV, respectively. The CR band in the T-shaped isomer is less intense and appears at the same energy as in the H-bonded dimer (0.11 eV). For all three isomers, the spectra change dramatically upon relaxation. In the stacked isomer, the intense CR band shifts to higher energies (i.e., from 0.5 to 1.3 eV) and becomes even more intense. In the H-bonded isomer, the CR bands (present at the neutral geometry at 0.1 eV) disappear upon proton transfer and the spectrum becomes very similar to that of the monomer. In the T-shaped isomer, new intense lines corresponding to charge-transfer transitions develop at 2.5−3.0 eV. Thus, the spectral evolution in these isomers is rather different, which may be exploited for their experimental determination.

Acknowledgment. We are grateful to Dr. Ksenia Bravaya for her insightful remarks and critical reading of the paper. This work was conducted in the framework of the iOpenShell Center for Computational Studies of Electronic Structure and Spectroscopy of Open-Shell and Electronically Excited Species (iopenshell.usc.edu) supported by the National Science Foundation through Grants CRIF:CRF CHE-0625419+0624602+0625237 and CHE-0616271.

Appendix: Performance of ωB97X-D for the Structures and Energetics of Noncovalent Neutral and Ionized Dimers

Self-interaction-corrected functionals provide more a reliable (although not fully satisfactory) description of the ionized noncovalent dimers than the standard functionals. To investigate the performance of the ωB97X-D functional[55] as an inexpensive alternative to more reliable wave function methods, we benchmarked this functional using the stacked uracil isomer. We compared the intra and interfragment structural parameters of the ωB97X-D/6-311(+)G**-optimized geometries of the neutral and the cation with the best available geometries. For the neutral system, the geometry from the S22 set of Hobza and co-workers was used as a benchmark.[59] For the cation, we compared against the IP-CISD/6-31(+)G*-optimized geometry. The average absolute errors and the standard deviations for the bond lengths and angles in the DFT-D-optimized geometries were calculated. In the neutral, the average absolute error and the standard

deviation for the bond lengths were 0.004 and 0.003 Å, respectively; the average absolute error and standard deviation for the angles were 0.247 and 0.182°. In the cation, the corresponding values were 0.010 and 0.005 Å and 0.377 and 0.233°. As for the interfragment parameters, in the neutral the DFT-D parameters (C_5-C_6 and O_2-N_1) differ by less than 0.05 Å from the geometry from the S22 set, while in the cation DFT-D overestimated them by 0.15 Å relative to the IP-CISD/6-31(+)G* value. Given the tendency of IP-CISD to overestimate the interfragment distances in weakly bound systems by 0.2−0.3 Å (as compared to the more accurate IP-CCSD),[51] the DFT-D geometry of the cation may be more accurate than the IP-CISD geometry. We conclude that the ωB97X-D structures are fairly accurate, which validates the use of this method for geometry optimizations of the ionized dimers.

To assess the performance of the ωB97X-D functional for the energetics, we computed the dissociation energies for all isomers of the neutral and cation dimers and compared them to the IP-CCSD/6-311(+)G** values. The results are summarized in Figures 4 and 5. ωB97X-D predicts the correct relative ordering of the neutral and cation isomers. Quantitatively, the DFT-D errors in dissociation energies with respect to the IP-CCSD values are in the 1−2 kcal/mol range for the neutral dimers and in the 1−5 kcal/mol range for the cations. The errors in D_e are nonsystematic. Therefore, DFT-D with the ωB97X-D functional provides a correct qualitative picture for the energetics; the quantitative predictions are of moderate accuracy, so a more reliable approach should be employed.

Supporting Information Available: Optimized geometries, corresponding reference energies, and frequencies (TXT). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Nácuteuñez, M. E.; Hall, D. B.; Barton, J. K. *Chem. Biol.* **1999**, *6*, 85.

(2) Henderson, P. T.; Jones, D.; Hampikian, G.; Kan, Y.; Schuster, G. B. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 8353.

(3) Lewis, F. D.; Letsinger, R. L.; Wasielewski, M. R. *Acc. Chem. Res.* **2001**, *34*, 159.

(4) Candeias, L. P.; Steenken, S. *J. Am. Chem. Soc.* **1989**, *111*, 1094.

(5) Hutter, M.; Clark, T. *J. Am. Chem. Soc.* **1996**, *118*, 7574–7577.

(6) Ghosh, A. K.; Schuster, G. B. *J. Am. Chem. Soc.* **2006**, *128*, 4172.

(7) de Vries, M. S. In *Radiation Induced Molecular Phenomena in Nucleic Acids*; Shukla, M., Leszczynski, J., Eds.; Springer: Berlin, 2008; p 323.

(8) Dougherty, D.; Wittel, K.; Meeks, J.; McGlynn, S. P. *J. Am. Chem. Soc.* **1976**, *98*, 3815.

(9) Urano, S.; Yang, X.; LeBrenton, P. R. *J. Mol. Struct.* **1989**, *214*, 315.

(10) Lauer, G.; Schäfer, W.; Schweig, A. *Tetrahedron Lett.* **1975**, *16*, 3939.

(11) Yu, C.; O'Donnel, T. J.; LeBreton, P. R. *J. Phys. Chem.* **1981**, *85*, 3851.

(12) Kim, S. K.; Lee, W.; Herschbach, D. R. *J. Phys. Chem.* **1996**, *100*, 7933.

(13) Satzger, H.; Townsend, D.; Stolow, A. *Chem. Phys. Lett.* **2006**, *430*, 144.

(14) Belau, L.; Wilson, K. R.; Leone, S. R.; Ahmed, M. *J. Phys. Chem. A* **2007**, *111*, 7562.

(15) Cauët, E.; Dehareng, D.; Liévin, J. *J. Phys. Chem. A* **2006**, *110*, 9200.

(16) Roca-Sanjuán, D.; Rubio, M.; Merchán, M.; Serrano-Andrés, L. *J. Chem. Phys.* **2006**, *125*, 084302.

(17) Cauët, E.; Liévin, J. *Adv. Quantum Chem.* **2007**, *52*, 121.

(18) Hudock, H. R.; Levine, B. G.; Thompson, A. L.; Satzger, H.; Townsend, D.; Gador, N.; Ulrich, S.; Stolow, A.; Martínez, T. J. *J. Phys. Chem. A* **2007**, *111*, 8500.

(19) Bravaya, K. B. ; Kostko, O. ; Ahmed, M. ; Krylov, A. I. *Phys. Chem. Chem. Phys.*, in press.

(20) Kostko, O.; Bravaya, K. B.; Krylov, A. I.; Ahmed, M. *Phys. Chem. Chem. Phys.*, submitted for publication.

(21) Golubeva, A. A.; Krylov, A. I. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1303.

(22) Colson, A.-O.; Besler, B.; Sevilla, M. D. *J. Phys. Chem.* **1992**, *96*, 9787.

(23) Colson, A.-O.; Besler, B.; Sevilla, M. D. *J. Phys. Chem.* **1993**, *97*, 13852.

(24) Sugiyama, H.; Saito, I. *J. Am. Chem. Soc.* **1996**, *118*, 7063.

(25) Prat, F.; Houk, K. N.; Foote, C. S. *J. Am. Chem. Soc.* **1998**, *120*, 845.

(26) Schumm, S.; Prévost, M.; Garcia-Fresnadillo, D.; Lentzen, O.; Moucheron, C.; Krisch-De Mesmaeker, A. *J. Phys. Chem. B* **2002**, *106*, 2763.

(27) Roca-Sanjuán, D.; Merchán, M.; Serrano-Andrés, L. *Chem. Phys.* **2008**, *349*, 188.

(28) Steenken, S. *Chem. Rev.* **1989**, *89*, 503.

(29) Kumar, A.; Sevilla, M. D. *J. Phys. Chem. B*, in press.

(30) Bertran, J.; Oliva, A.; Rodríguez-Santiago, L.; Sodupe, M. *J. Am. Chem. Soc.* **1998**, *120*, 8159.

(31) Mulliken, R. S. ; Person, W. B. *Molecular Complexes*; Wiley-Interscience: New York, 1969.

(32) Badger, B.; Brocklehurst, B. *Trans. Faraday Soc.* **1969**, *65*, 2576.

(33) Badger, B.; Brocklehurst, B. *Trans. Faraday Soc.* **1969**, *65*, 2582.

(34) Badger, B.; Brocklehurst, B. *Trans. Faraday Soc.* **1969**, *65*, 2588.

(35) Pieniazek, P. A.; Krylov, A. I.; Bradforth, S. E. *J. Chem. Phys.* **2007**, *127*, 044317.

(36) Pieniazek, P. A.; Bradforth, S. E.; Krylov, A. I. *J. Chem. Phys.* **2008**, *129*, 074104.

(37) Pieniazek, P. A.; VandeVondele, J.; Jungwirth, P.; Krylov, A. I.; Bradforth, S. E. *J. Phys. Chem. A* **2008**, *112*, 6159.

(38) Pieniazek, P. A.; Sundstrom, E. J.; Bradforth, S. E.; Krylov, A. I. *J. Phys. Chem. A* **2009**, *113*, 4423.

(39) Müller-Dethlefs, K.; Hobza, P. *Chem. Rev.* **2000**, *100*, 143.

(40) Sponer, J.; Leszczynski, J.; Hobza, P. *Biopolymers* **2002**, *61*, 3.

(41) Saigusa, H. *Photochem. Photobiol.* **2006**, *7*, 197.

(42) de Vries, M. S.; Hobza, P. *Annu. Rev. Phys. Chem.* **2007**, *58*, 585.

(43) Davidson, E. R.; Borden, W. T. *J. Phys. Chem.* **1983**, *87*, 4783.

(44) Russ, N. J.; Crawford, T. D.; Tschumper, G. S. *J. Chem. Phys.* **2004**, *120*, 7298.

(45) Polo, V.; Kraka, E.; Cremer, D. *Mol. Phys.* **2002**, *100*, 1771.

(46) Zhang, Y.; Yang, W. *J. Chem. Phys.* **1998**, *109*, 2604.

(47) Sinha, D.; Mukhopadhyay, D.; Mukherjee, D. *Chem. Phys. Lett.* **1986**, *129*, 369.

(48) Pal, S.; Rittby, M.; Bartlett, R. J.; Sinha, D.; Mukherjee, D. *Chem. Phys. Lett.* **1987**, *137*, 273.

(49) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1994**, *101*, 8938.

(50) Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1995**, *102*, 3629.

(51) Golubeva, A. A.; Pieniazek, P. A.; Krylov, A. I. *J. Chem. Phys.* **2009**, *130*, 124113.

(52) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540.

(53) Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.

(54) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

(55) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(56) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.

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

(58) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(59) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(60) Olaso-Gonzáles, G.; Roca-Sanjuán, D.; Serrano-Andrés, L.; Merchán, M. *J. Chem. Phys.* **2006**, *125*, 231002.

(61) Bravaya, K. Private communication.

(62) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *Comput. Phys. Commun.* **2008**, *9*, 1636.

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

(64) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656.

(65) Kratochvil, M.; Engkvist, O.; Sponer, J.; Jungwirth, P.; Hobza, P. *J. Phys. Chem. A* **1998**, *102*, 6921.

CT900515A

JCTC Journal of Chemical Theory and Computation

Blending Determinism with Evolutionary Computing: Applications to the Calculation of the Molecular Electronic Structure of Polythiophene

Kanchan Sarkar, Rahul Sharma, and S. P. Bhattacharyya*

Department of Physical Chemistry, Indian Association for the Cultivation of Science, Jadavpur, Kolkata-700032, India

Received October 13, 2009

Abstract: A density matrix based soft-computing solution to the quantum mechanical problem of computing the molecular electronic structure of fairly long polythiophene (PT) chains is proposed. The soft-computing solution is based on a "random mutation hill climbing" scheme which is modified by blending it with a deterministic method based on a trial single-particle density matrix [$\mathbf{P}^{(0)}(R)$] for the guessed structural parameters (R), which is allowed to evolve under a unitary transformation generated by the Hamiltonian $\mathbf{H}(R)$. The Hamiltonian itself changes as the geometrical parameters (R) defining the polythiophene chain undergo mutation. The scale (λ) of the transformation is optimized by making the energy [$E(\lambda)$] stationary with respect to λ. The robustness and the performance levels of variants of the *algorithm* are analyzed and compared with those of other derivative free methods. The method is further tested successfully with optimization of the geometry of bipolaron-doped long PT chains.

1. Introduction

Exact solutions of the Schrodinger equation for many-electron systems are impossible to obtain analytically. The solutions are therefore obtained by various approximation methods, which are still rather complicated for applications to large systems. With rapid and spectacular advances in digital computing, computational methods of electronic structure calculations have attracted serious attention in recent years.[1−3] These methods have become practically essential for understanding large systems at the microscopic level. Finding efficient algorithms for handling large quantum systems even at an approximate level has become a challenging issue. Traditional electronic structure algorithms calculate eigenstates associated with discrete energy levels, and this leads to a diagonalization problem of the Hamiltonian matrix. The computational effort in traditional diagonalization methods scales as N^3, where N is the dimension of the basis space. Usually one is interested in finding the lowest energy structure so that one has to simultaneously search the potential energy surface $E(R)$ for locating the global minimum while diagonalizing the Hamiltonian matrix $\mathbf{H}(R)$ at different geometries (R).

In the present paper we have proposed a novel nondeterministic *algorithm* for locating the minimum energy structures of neutral or doped PT chains of 100, 150, and 200 thiophene rings. To be specific, we have proposed a variant of the "directed random mutation hill climbing (DRMHC) method",[14,15] which works with a randomly generated string of all the geometrical parameters (nuclear position variables, R) required to compute the energy and therefore the fitness of a neutral or doped PT molecule within the framework of a modified SSH effective π-electron Hamiltonian model.[12,13] The geometry string $\{R_i\}$ is allowed to undergo directed random mutations. The string of mutated geometry variables $\{R_i'\}$ is used to define the Hamiltonian $\mathbf{H}(\{R_i'\})$, which acts as the generator of a unitary transformation, $\mathbf{U}(\lambda,\{R_i'\})$. $\mathbf{U}(\lambda,\{R_i'\})$ transforms a trial one-electron density matrix ($\mathbf{P}^{(0)}$) into a "mutated" one-electron density matrix ($\mathbf{P}^{(0)\prime}$), which is used to compute the "energy" of the mutated structure and hence the fitness of the structure coded by the mutated geometry string. The parameter λ fixes the scale of the transformation which is optimized at each random mutation hill climbing step by making the energy stationary with

* Corresponding author fax: (91)332473 2805; e-mail: pcspb@iacs.res.in.

10.1021/ct900540d
Published on Web 02/01/2010

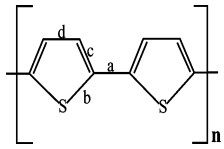


Figure 1. Fragment of the thiophene chain: (a) C−C bridging bond, (b) C−S single bond, (c) C−C double bond, (d) C−C single bond.

Table 1. Parameters in the SSH Hamiltonian (for Polythiophene) Used in These Calculations

param	value	description
A	123.6 eV	Hamiltonian parameter
B	0.3776 au	Hamiltonian parameter
D	7.814 au^{-1}	Hamiltonian parameter
$R^0_{(C-C)}$	1.557 au	C−C single bond length
$R^0_{(C-S)}$	1.782 au	C−S bond length
$R^0_{(C-C)B}$	1.557 au	C−C bridging bond length

respect to λ. We note here that there have been a number of earlier attempts to calculate the one- and two particle-density matrices using similar unitary transformations[8−10] on a trial density. We have, to the best of our knowledge, tried for the first time to blend it with an otherwise stochastic approach for computing the one-electron density matrix. We have compared the λ-optimized DRMHC algorithm with several alternative soft-computing methods of solving the problem. The algorithm is elaborately tested on long polythiophene chains, undoped as well as doped, and the results are briefly analyzed.

The outline of the paper is as follows. In section 2 we present the *algorithm* in detail, while the results of the applications are described in section 3. Section 4 presents the concluding remarks.

2. Method

Polythiophene (PT) oligomers have attracted the widespread attention of experimentalists and theoreticians because these molecules and their derivatives are capable of displaying metallic conduction under electron or hole doping.[4−7,11−13] Neutral PTs are chemically stable, can be synthesized easily, and can be doped with dopants such as ClO_4 and AsF_5. The stability remains intact even after doping.

The UV photoelectron spectrum[5] of neutral PT shows the existence of two π-bands in the system, and the Fermi level is ∼1.2 eV above the valence band (VB). There are two nondegenerate classical resonating forms of PT in which the $2p_z$ orbitals of the carbon atoms and the $3p_z$ orbitals of the sulfur atoms interact to form a π-band, half of which is occupied by the electrons. Had the two forms been isoenergic, a solitonic mode of conduction would have been viable as the solitons have a tendency to separate because the structure between the defects is isoenergic with the structure outside. However, in PT, the degeneracy is weakly lifted, and as a consequence the solitonic mode of conduction is ruled out. It turns out that polarons and bipolarons are the most important excitations and charge-storage configurations in doped PT.[6]

While PT is a semiconductor with a band gap of about 2 eV, hole doping reduces the gap, and at high doping levels PT begins to show metallic properties. It is important therefore to understand how the electronic structure of neutral PT evolves under doping. It is expedient in this context to have algorithms that can locate the global minimum energy structures on the potential energy surface of PT oligomers, undoped or doped. There are several deterministic methodologies, mostly based on a gradient search, which have been explored for locating the minimum energy structures of PTs.[6,7] Rarely, however, nondeterministic techniques have been explored in this context.[11] Our primary concern here is to introduce such a hybrid technique of geometry optimization and calculation of electronic structure and test its workability with undoped and doped PTs as examples.

PT is treated as a conjugated polyene and described by a tight binding model of the π-electronic system that includes only the nearest neighbor hopping interactions. The lattice dynamics is treated classically. This leads to a modified SSH effective π-electronic Hamiltonian $\mathbf{H}$ where[12,13]

$$\mathbf{H} = \mathbf{H}_e^{\pi} + \mathbf{H}_l^{\sigma} \tag{1}$$

The π-electronic Hamiltonian ($\mathbf{H}_e^{\pi}$) is expressed as

$$\mathbf{H}_e^{\pi} = \sum_n \alpha_n a_n^{\dagger} a_n + \frac{1}{2}\sum_{n,m} V_{nm} a_n^{\dagger} a_m + hc \tag{2}$$

and the lattice Hamiltonian (the σ-electronic framework Hamiltonian) is defined as

$$\mathbf{H}_l^{\sigma} = \sum_n \frac{p_n^{\,2}}{2M_n} + \frac{1}{2}\sum_{n,m} f(R_{nm}) \tag{3}$$

α_n is the self-energy of the carbon $2p_z$ electron ($3p_z$ electron of sulfur) at the nth site, and V_{nm} represents the hopping interaction between the nth and mth sites. V_{nm} is parametrized in the form[7,11]

$$\begin{aligned} V_{nm} &= -Ae^{-R_{nm}/B} && (|n-m| = 1) \\ &= 0 && (\text{otherwise}) \end{aligned} \tag{4}$$

with two parameters, A and B. R_{nm} is the distance separating the nth and mth adjacent sites. We use frozen core approximation so that the kinetic energy of the lattice is zero, i.e.

$$\mathbf{H}_l^{\sigma} = \frac{1}{2}\sum_{n,m} f(R_{nm}) \tag{5}$$

where

$$f(R_{nm}) = -AD_{nm}(R_{nm} - R_{nm}^0 + B)e^{-R_{nm}/B}$$

A, B, and D_{nm} are parameters of the model and R_{nm}^0 is the standard length of the $n-m$ bond.[7,11]

We start by guessing randomly the M number of bond lengths required to describe the N_r-ring chain, forming and diagonalizing the N-dimensional ($N = 5N_r$) π-electron Hamiltonian $\mathbf{H}_e^{\pi}$ in the N-dimensional basis of the carbon $2p_z$ and sulfur $3p_z$ atomic orbitals of all the carbon and sulfur atoms of the chain. The diagonalization generates a set of N π molecular orbitals $\{\phi_i\}$, N_{oc} of which are occupied by the

π-electrons in pairs ($N_{oc} = 3N_r$). In general, we can write the π MOs $\{\phi_i\}$ as linear combinations of the atomic basis orbitals (χ_p), where

$$\phi_i = \sum_{p=1}^{N} c_{p_i}\chi_p$$
$$\mathbf{H}_e^{\pi}\phi_i = \varepsilon_i^{\pi}\phi_i,\ i = 1, 2, ..., N \tag{6}$$

ε_i^{π} represent the binding energies of the π molecular electrons. The linear expansion coefficients $\{c_{p_i}\}$ define the elements of the charge density bond order matrix $\mathbf{P}$ through the relation

$$\mathbf{P}_{pq} = \sum_{i=1}^{N_{oc}} c_{p_i} c_{q_i}^{*} \tag{7}$$

so that the starting density matrix is defined as

$$\mathbf{P}^{(0)} = \sum_{i=1}^{N_{oc}} c_i^0 c_i^{0\dagger} \tag{8}$$

$\mathbf{P}^{(0)}$ satisfies the following constraints:

$$\mathbf{P}^{(0)\dagger} = \mathbf{P}^{(0)},$$
$$2\ \mathrm{Tr}\,\mathbf{P}^{(0)} = N_e, \quad (N_e = 6N_r)$$
$$(\mathbf{P}^{(0)})^2 = \mathbf{P}^{(0)} \tag{9}$$

where N_e is the number of electrons. The ground-state π-electronic energy E_G^{π} is given by

$$E_G^{\pi} = \sum_{k=1}^{N_{oc}} n_k \varepsilon_k^{\pi} = 2\sum_{n=1} \alpha_n q_n^0 + 2\sum_{n<m}\sum_{m} V_{nm}\mathbf{P}_{nm}^0 \tag{10}$$

q_n^0 is the electronic charge density at the nth site (diagonal elements of $\mathbf{P}^{(0)}$ formed in the atomic orbital basis). The total energy of the π-electron system including the elastic deformation energy of the σ-electronic framework is given by

$$E_T(R) = E_G^{\pi} + E_{\sigma} = E_G^{\pi} + \frac{1}{2}\sum_{n,m} f(R_{nm}) \tag{11}$$

Our purpose is to find the set of $\{R_{nm}\}$ values that globally minimize the ground-state energy $E_T(R)$ and calculate the corresponding charge density bond order matrix ($\mathbf{P}$) that encodes the equilibrium charge distribution in the ground state. We suggest the following strategy for solving the problem.

Let the geometrical parameters $\{R_{nm}\}$ defining the PT chain form a string $s(R_1, R_2, R_3, ..., R_k, ..., R_m, ..., R_M)$, where M is the total number of bonds in the system. The π-electronic Hamiltonian $\mathbf{H}_e^{\pi}(R)$ for the guessed geometry coded by the string $s(R)$ is diagonalized in the basis of the $2p_z$ orbitals of the carbon atoms and $3p_z$ orbitals of the sulfur atoms, and a trial density $\mathbf{P}^{(0)}(R)$ is generated. We choose one of the parameters, say the kth parameter, randomly with a

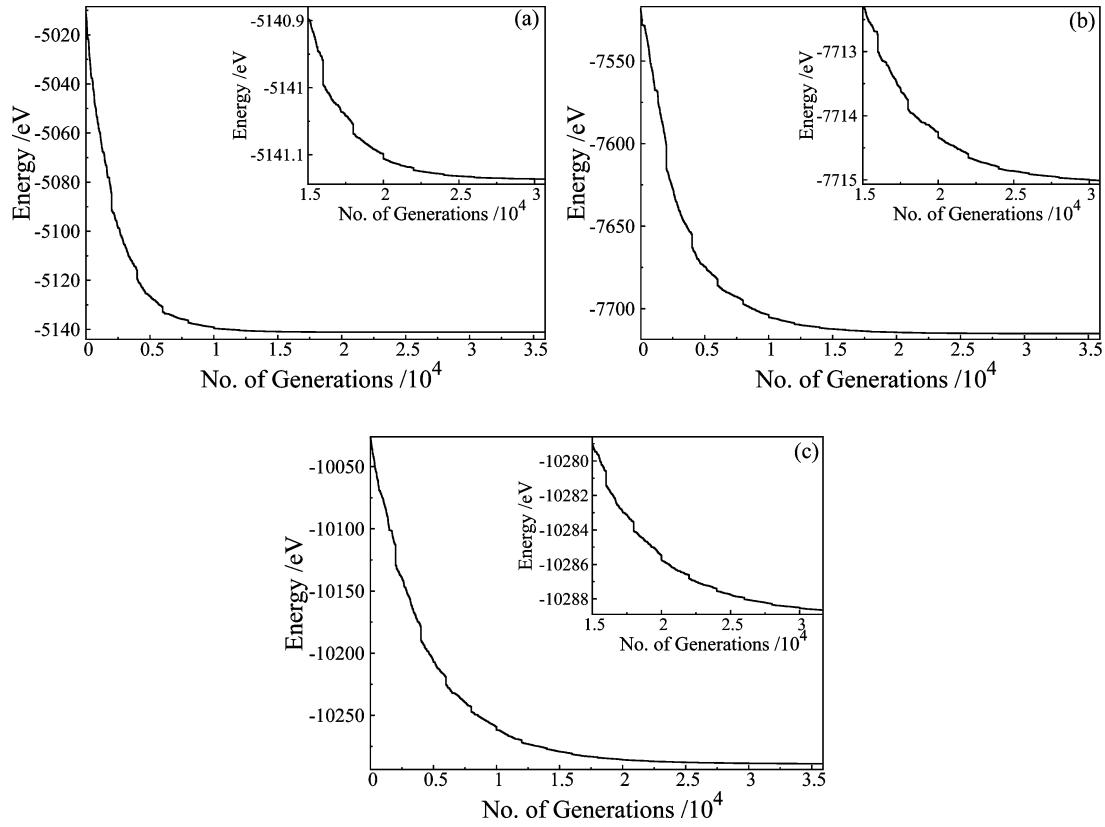


Figure 2. Energy evolution profiles of PT chains in the λ-optimized density matrix based DRMHC method (the inset figures display the fitness evolution between 15000 and 30000 generations in all the cases of 100-, 150-, and 200-ring PT chains): (a) for the 100-ring chain, (b) for the 150-ring chain, (c) for the 200-ring chain.

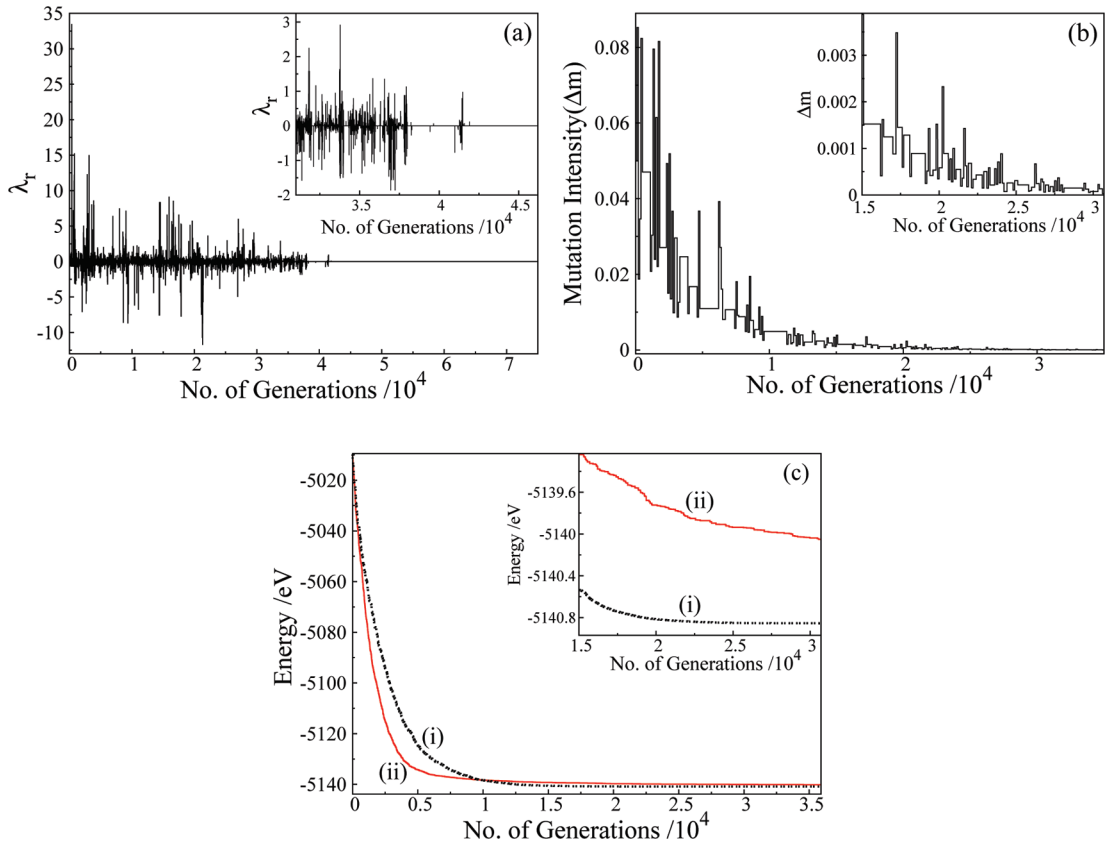


Figure 3. For a PT chain containing 100 rings: (a) evolution of λ_r in the λ-optimized density matrix based DRMHC scheme (the inset figure displays the evolution of λ_r between 30000 and 45000 generations), (b) evolution of Δ_m in the DRMHC scheme, (c) (i) energy evolution profiles in the DRMHC method at fixed $\lambda = 0.1$ (in the inset the evolution between 15000 and 30000 generations is shown to emphasize the different rates of the evolution in the convergence region), (ii) energy evolution profiles in the RMHC method at fixed $\Delta_m = 0.05$ (the inset shows the evolution profile between 15000 and 30000 generations).

probability p_m for mutation. The mutation induces a small change in the chosen parameter (R_k) in the following manner:[16]

$$R'_k = R_k + (-1)^l \Delta_m r$$

where l is a random integer, r is a random number in the range (0, 1), and Δ_m is a directed mutation intensity[17] (see later). The mutated string $s(R_1, R_2, R_3, ..., R'_k, ..., R_m, ..., R_M)$ is used to generate the π-electron Hamiltonian $\mathbf{H}_e^\pi(R')$, which in turn generates the unitary transformation matrix

$$\mathbf{U}_\lambda = e^{-i\lambda \mathbf{H}_e^\pi(R')} \tag{12}$$

where λ defines the scale of the transformation. The mutated density matrix $\mathbf{P}^{(1)}(R')$ is generated by transforming $\mathbf{P}^{(0)}(R)$ with $\mathbf{U}_\lambda$ in the following manner:[18]

$$\begin{aligned}\mathbf{P}^{(1)}(R') &= \mathbf{U}_\lambda \mathbf{P}^{(0)}(R) \mathbf{U}_\lambda^\dagger \\ &= e^{-i\lambda \mathbf{H}_e^\pi(R')}\, \mathbf{P}^{(0)}(R)\, e^{i\lambda \mathbf{H}_e^\pi(R')}\end{aligned} \tag{13}$$

Expanding the exponentials in powers of λ, we have (up to second order in λ)

$$\mathbf{P}^{(1)}(R') \approx \mathbf{P}^{(0)}(R) - i\lambda[\mathbf{P}^{(0)}(R), \mathbf{H}_e^\pi(R')] - \frac{\lambda^2}{2!}[[\mathbf{P}^{(0)}(R), \mathbf{H}_e^\pi(R')], \mathbf{H}_e^\pi(R')] \tag{14}$$

In eq 14, [**A**, **B**] represents the commutator of matrices **A** and **B**. The problem now is to fix the scale parameter λ. We do that by estimating the electronic energy $\varepsilon_{el}^\pi(\lambda)$ and making it stationary with respect to variations in λ. Thus, using $\mathbf{P}^{(1)}$ to estimate $\varepsilon_{el}^\pi(\lambda)$, we have, up to second order in λ

$$\begin{aligned}\varepsilon_{el}^\pi(\lambda) &\approx 2\,\mathrm{Tr}\{\mathbf{P}^{(1)}\,\mathbf{H}_e^\pi(R')\} \\ &= 2[\mathrm{Tr}\{\mathbf{P}^{(0)}\,\mathbf{H}_e^\pi(R')\}] - i\lambda\,\mathrm{Tr}\{[\mathbf{P}^{(0)}, \mathbf{H}_e^\pi(R')]\mathbf{H}_e^\pi(R')\} - \frac{\lambda^2}{2!}\mathrm{Tr}\{[[\mathbf{P}^{(0)}, \mathbf{H}_e^\pi(R')], \mathbf{H}_e^\pi(R')]\mathbf{H}_e^\pi(R')\}\end{aligned}$$

Setting

$$\frac{\partial \varepsilon_e^\pi(\lambda)}{\partial \lambda} = 0 \tag{15}$$

then leads to

$$\lambda_{opt} = -i\frac{\mathrm{Tr}\{[\mathbf{P}^{(0)}, \mathbf{H}_e^\pi(R')]\mathbf{H}_e^\pi(R')\}}{\mathrm{Tr}\{[[\mathbf{P}^{(0)}, \mathbf{H}_e^\pi(R')], \mathbf{H}_e^\pi(R')]\mathbf{H}_e^\pi(R')\}} \tag{16}$$

$$= -i\lambda_r \tag{17}$$

where

$$\lambda_r = \frac{\mathrm{Tr}\{[\mathbf{P}^{(0)}, \mathbf{H}_e^\pi(R')]\mathbf{H}_e^\pi(R')\}}{\mathrm{Tr}\{[[\mathbf{P}^{(0)}, \mathbf{H}_e^\pi(R')], \mathbf{H}_e^\pi(R')]\mathbf{H}_e^\pi(R')\}} \tag{18}$$

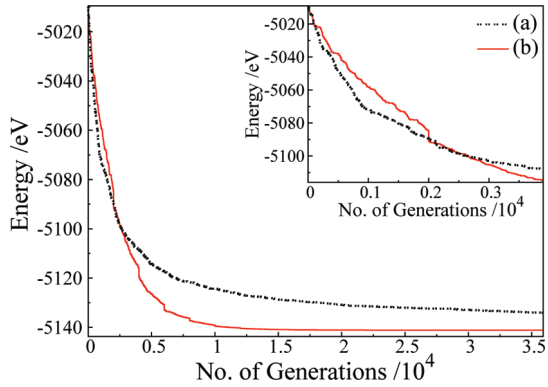


Figure 4. Comparison between the repeated diagonalization based RMHC procedure in the case of a 100-ring PT chain (a) and the λ-optimized density matrix based DRMHC method (b) (the inset shows that up to 3000 generations the two schemes perform in more or less the same way).

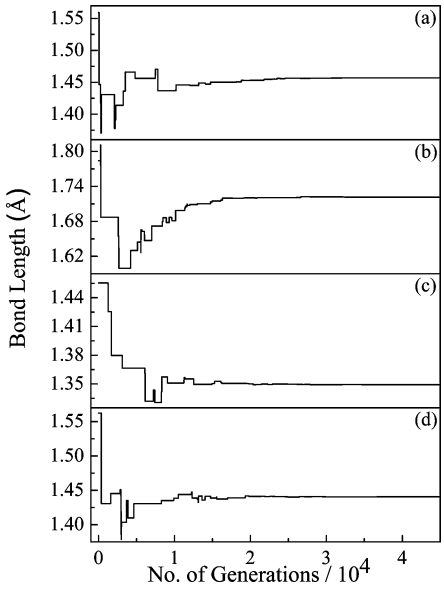


Figure 5. Bond length evolution profiles of 100-ring PT chains in the λ-optimized DRMHC method: (a) C−C bridging bond, (b) C−S single bond, (c) C−C double bond, (d) C−C single bond.

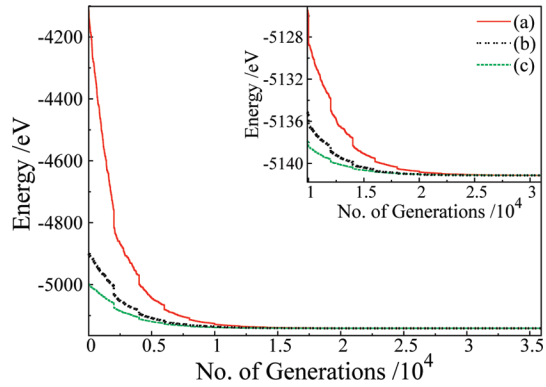


Figure 6. Convergence profiles with different starting points suggest that the λ-optimized density matrix based DRMHC process is robust (the inset shows profiles in the region between 9900 and 27000 generations). The molecule studied is a 100-ring PT chain.

If the total energy of the mutated structure $E'_T(R') = \varepsilon'^{\pi}_e + \varepsilon'_\sigma$ computed with λ_{opt} is lower (or the absolute value of $E'_T(R')$, which we call fitness, is higher) than the premutation energy $E_T(R)$, the mutated geometry string $s\{R'_i\}$ replaces the old string $s\{R_i\}$. If the condition is not satisfied, a new mutation site is chosen on the old string; thus, the algorithm avoids evolving nonviable candidates. In either case, one generation is counted to have elapsed. We may note here that $\mathbf{P}^{(1)}(R')$ of eq 13 is strictly idempotent while $\mathbf{P}^{(1)}(R')$ of eq 14 may not be so because of truncation. Also 2 Tr $\mathbf{P}^{(1)}(R')$ may deviate from the total number of electrons. Therefore, these quantities are followed throughout the evolution and corrective measures taken if required (through standard trace purification or idempotency-correcting routines).

The directed mutation scheme used in the present case[17] uses a mutation intensity that is dynamically adjusted on the basis of the degree of acceptability of mutation over a number of past generations. Thus, if the number of accepted mutations in the last 100 generations is less than 10, Δ_m is lowered to $\Delta'_m = \Delta_m/(1 + r)$ (r is a random number between 0 and 1). It is enhanced to $\Delta'_m = \Delta_m/(1 - r)$ if the number of accepted mutations is greater than 20 in the last 100 generations. Δ_m is kept unchanged otherwise.

The process is continued until the total energy stops evolving further. The equilibrium ground-state structure is represented by the geometry variables that the string $s(R_1, R_2, R_3, ..., R_M)$ corresponding to the converged energy or fitness encodes. The same algorithm can be extended to determine the electronic structure of doped PTs or other similar molecules.

3. Results and Discussion

We have considered specific cases of polythiophene oligomers containing 100, 150, or even 200 rings in a chain which require 599, 899, and 1199 bond lengths or geometry variables, respectively, to define the SSH Hamiltonian.[12,13] The geometry strings $s(R)$ encoding the structures of PT oligomers are arrays containing 599, 899, and 1199 floating point variables, respectively. We have experimented with several variants of the *algorithm* proposed in section 2, a few of which are reported here. The parameters of the SSH Hamiltonian used in these calculations are reported in Table 1.

3.1. Geometry Optimization in Neutral Polythiophene Oligomers. Let us consider a 100-ring undoped PT chain. In our λ-optimized DRMHC algorithm, we start with a geometry string in which the geometry variables are allowed to be distributed randomly over a predefined range. The mutation probability p_m is set to have a value $p_m = 5/M$, where M is the total number of bonds in the system, and held fixed throughout the evolution. The initial mutation intensity or rate (Δ_m) is assumed to be 0.05. Empirically Schaffer et al.[19] found that the optimum mutation rate in a genetic algorithm (GA) can be represented by the formula

$$\ln(N) + 0.93 \ln(\Delta_m) + 0.45 \ln(n) = 0.56 \qquad (19)$$

where N = size of the population, Δ_m = mutation rate, and n = length of the chromosome. Hesser et al.[20] provided a

heuristic argument in support of the empirical formula. The initial value of Δ_m chosen by us happens to be around half of the empirically estimated value of Δ_m (optimum) predicted by eq 19 if we assume that n represents the number of parameters present in a geometry string used here. The π-electronic Hamiltonian ($\mathbf{H}_e^{\pi}$) is diagonalized once after every 2000 generations. The parameter λ of the unitary transformation is optimized as outlined in section 2 (eq 16).

Figure 1 shows various bonds in the PT chain. Parts a−c of Figure 2 show how energy evolves during the density matrix based λ-optimized DRMHC search for the thiophene chain containing 100, 150, and 200 rings, respectively. In about 15000 generations, the search locates the gross features of the equilibrium structure in each case. The typical evolution pattern of λ_r (eq 18) for the representative case of λ optimization in the case of a 100-ring neutral PT chain is profiled in Figure 3a.

There are wide fluctuations in the scale parameter λ_r from one generation to another. As the search converges to the equilibrium structure the fluctuations get strongly damped and $\lambda_r \rightarrow 0$. Far away from the global minimum large λ_r is predicted to have a large magnitude. In a sense, this amounts to emphasizing the breadth search aspect of the algorithm. As the search approaches the global minimum, $\lambda_r \rightarrow 0$, thereby emphasizing the depth search aspect. The corresponding evolution profile of directed mutation intensities is depicted in Figure 3b. To understand the role played by λ-optimization, we have reported a second set of calculations on 100-ring PT chains in which λ_r is kept fixed at $\lambda_{av} = 0.1$ while all other parameters of the algorithm remain unaltered. The search failed to converge to the desired level, indicating the importance of λ-optimization in conjuction with a DRMHC search (Figure 3c). Similarly, another set of calculations were carried out with fixed mutation intensity ($\Delta_m = 0.05$) to understand the role played by directed mutation (Figure 3d). Again the search failed to hit the global minimum energy structure. We have checked that the trace of the density matrix and its idempotent character remain conserved to the desired accuracy throughout the evolution. Therefore, additional computational labor for purifying the "trace" or restoring the idempotency of $\mathbf{P}^{(1)}(R')$ of eq 14 was avoided. The transformation shown in eq 13 thus remains unitary for all practical purposes after the truncation shown in eq 14.

The performance of the λ-optimized DRMHC technique has also been compared with that of the conventional repeated diagonalization based procedure. Here $\mathbf{H}_e^{\pi}$ was diagonalized following every successful mutation step, and a new $\mathbf{P}^{(0)}$ was calculated directly from the eigenvectors of $\mathbf{H}_e^{\pi}$, bypassing the construction of and transformation with the unitary matrix $\mathbf{U}(\lambda,\{R_i\})$. It can be seen that the repeated diagonalization based RMHC procedure (Figure 4a) and density matrix based λ-optimized DRMHC (Figure 4b) algorithm perform more or less identically in the initial phase of evolution (first 3000 generations or so). Beyond that point density matrix based λ-optimized DRMHC performs better (Figure 4). We may mention that the maximum number of diagonalizations required in density matrix based λ-optimized DRMHC is between 10 and 20, while the repeated diagonalization RMHC procedure requires between 40000 and 45000 diagonalizations. Even then, the repeated diagonalization based RMHC procedure fails to hit the global minimum energy structure. The results are summarized in Table 2.

Parts a−d of Figure 5 display how the different geometry variables evolve during a search in the density matrix based λ-optimized DRMHC scheme for a PT chain containing 100

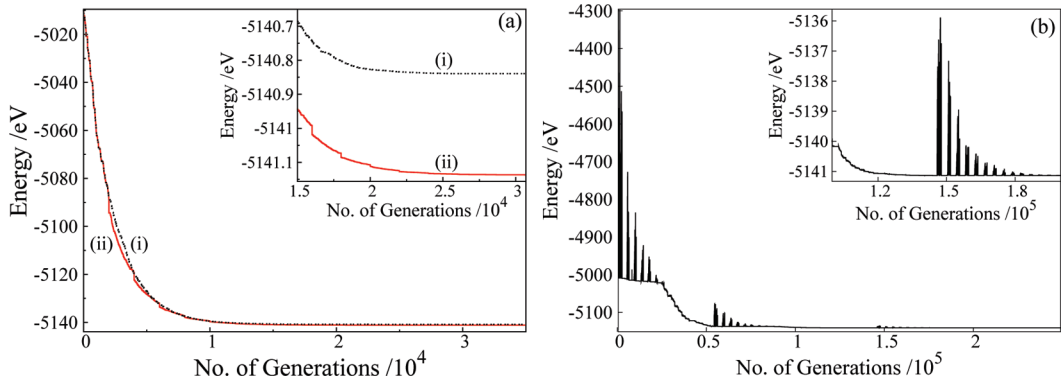


Figure 7. Energy evolution profiles of 100-ring PT chains with (a) (i) the GA (the inset shows the profile in the region between 15000 and 30000 generations) and (ii) the restricted crossover GA (crossover operation is restricted up to 2500 generations) (the inset shows the profile in the region between 15000 and 30000 generations) and (b) simulated annealing, with the inset representing the profile in the region between 99900 and 199900 generations.

Table 2. Comparison of the λ-Optimized Density Matrix Based DRMHC Method with the Repeated Diagonalization Based RMHC Procedure and Other Derivatives of the Density Matrix Based RMHC Scheme

scheme used	energy (eV)	no. of generations required	comment
λ-optimized DRMHC	−5141.137	37565	smooth and fast
DRMHC with fixed λ ($\lambda = 0.1$)	−5140.855	65201	gets stuck close to the global minimum
RMHC with fixed mutation intensity ($\Delta_m = 0.05$)	−5140.320	49375	fails to find the global minimum structure
repeated diagonalization	−5135.216	44752	fails to converge to the global minimum

Table 3. Comparison of the Performance of the λ-Optimized Density Matrix Based DRMHC Method and Other Soft-Computing Methods

scheme used	energy (eV)	no. of energy evaluations or iteration steps required	comment
simulated annealing method	−5141.137	240000	
genetic algorithm	−5141.137	67638	
genetic algorithm with restricted crossover	−5141.137	35539	costly but converges

rings. There are wide fluctuations in the bond lengths in the initial phase, which gradually get damped as the search converges.

To test the robustness and stability of the algorithm, the calculations have been performed with widely different starting geometries on the potential energy surface. The performance was slightly worse or better in different runs with different inputs (Figure 6), but convergence was achieved in each case. The average performance level closely matches with what has been reported in this paper.

3.2. Comparison with Other Soft-Computing Methods. The energy evolution of the density matrix based λ-optimized DRMHC method has been compared with what one actually observes with a density matrix based GA with a population size of 4 and the density matrix based simulated annealing method (SAM).[18] The convergence profiles of fitness/energy obtained with the GA and SAM are displayed in parts a and c, respectively, of Figure 7 for the geometry optimization of a 100-ring neutral PT chain. We note that the λ-optimized DRMHC method outperforms the two other soft-computing methods in terms of the number of steps needed to reach the minimum energy structure. In the case of the GA, the crossover operation is expected to help in the achievement of faster convergence, which is indeed noticed in the initial stages; however, the convergence to the global minimum is delayed. We have also compared the performance of a GA with restricted crossover (crossover allowed initially up to a limited number of generations) with that of the density matrix based λ-optimized DRMHC procedure (Figure 7b). It is observed that the restricted crossover GA provides approximately the same performance as obtained with the λ-optimized DRMHC scheme. DRMHC is nevertheless computationally economic as it needs only one string evaluation per generation, whereas in the GA fitness evaluation of a population of strings (four here) is to be carried out in every generation. The performance statistics are reported in Table 3. The scale optimized density matrix based DRMHC method has also been used to compute the global minimum energy structures of neutral PT chains of 150 and 200 rings. The predicted equilibrium geometrical parameters are reported in Table 4. The predicted lengths are practically identical. We did not enforce any symmetry constraint on the chain. Every geometrical parameter was allowed to evolve freely. The symmetry appeared naturally as the search converged to the global minimum.

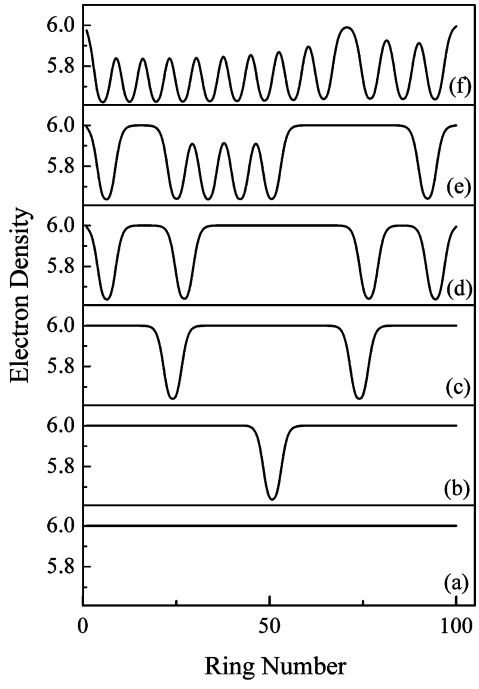


Figure 8. Density of electrons in the different rings of the 100-ring PT chain for bipolaron-doped structures: (a) 0-bipolaron-, (b) 1-bipolaron-, (c) 2-bipolaron-, (d) 4-bipolaron-, (e) 6-bipolaron-, and (f) 12-bipolaron-doped PT chains.

3.3. Geometry Optimization in Doped Polythiophene Oligomers. Doping of PTs creates structurally favored polaronic and bipolaronic defect states responsible for charge storage and excitation. The distortion energies (E_{dis}) of formation for two polarons and one bipolaron are the same, but the decrease in ionization energy is much more pronounced in the case of bipolarons ($2\Delta\varepsilon^{bip}$) than for two polarons ($2\Delta\varepsilon^{pol}$), making one bipolaron more stable than two polarons in these systems.[21]

We have investigated the evolution of structures of bipolaron-doped 100- and 150-ring PT chains containing up to 12 bipolaronic defects. Since there is a large-scale reorganization of geometry following doping, the doped systems provide challenging examples for testing the power of the new technique of geometry optimization introduced here. Let us first inspect what happens when a bipolaronic defect is incorporated into a 100-ring PT chain. Following the creation of a defect (removal of the electron pair from the HOMO), the chain is allowed to relax and the geometrical parameters are globally optimized by using λ-optimized DRMHC in the way mentioned in section 2. The starting geometry is assumed to be identical with the neutral $(PT)_{n=100}$ geometry. The relaxation causes the 2 units of positive charge to spread over the constituent rings of the chain. The charge distributions in neutral and one-bipolaron-doped PT chains of 100 rings are compared in Figure 8a,b. The maximum positive charge is seen to accumulate at the middle of the chain and get distributed on the two sides with near perfect Gaussian symmetry (Figure 8b). When a second bipolaronic defect is incorporated into the 100-ring PT chain, the charge distribution has a two-peak symmetric structure, as if the two defects tend to get separated and localized in two different regions of the chain (Figure 8c). The symmetry is destroyed as more bipolarons are introduced (Figure 8d−f). The one-bipolaron-doped 100-ring PT chain has four clearly

Table 4. Optimized Bond Lengths in 100-, 150-, and 200-Ring Neutral PT Chains[a]

no. of rings	optimized C−S bond length (Å)	optimized C−C bond length (Å)	optimized C=C bond length (Å)	optimized C−C junction bond (Å)
100	1.72152 (1.72153)	1.4405 (1.4406)	1.3492 (1.3493)	1.4569 (1.457)
150	1.72152 (1.72153)	1.4405 (1.4406)	1.3492 (1.3493)	1.4569 (1.457)
200	1.72152 (1.72153)	1.4405 (1.4406)	1.3492 (1.3493)	1.4569 (1.457)

[a] Refer to Figure 1.

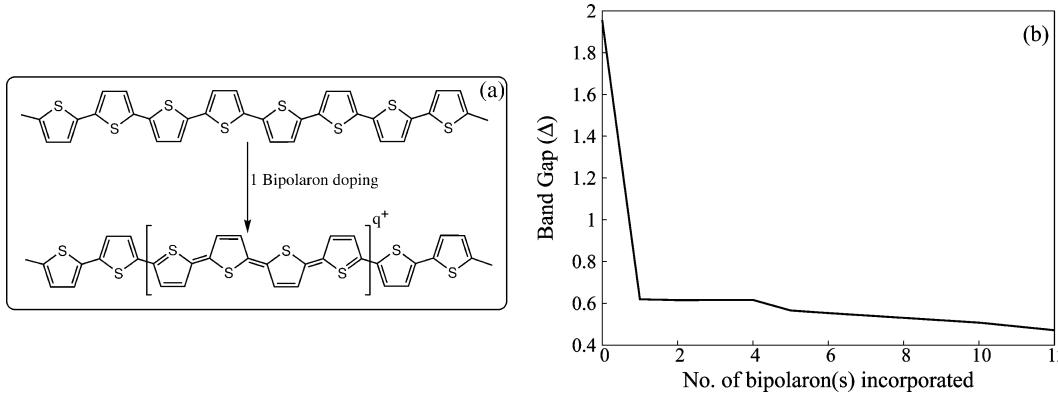


Figure 9. (a) Chemical structure of a one-bipolaron-doped 100-ring PT chain ($q = 1.36$). (b) Reduction of the band gap with the number of bipolarons incorporated into the 100-ring PT chain.

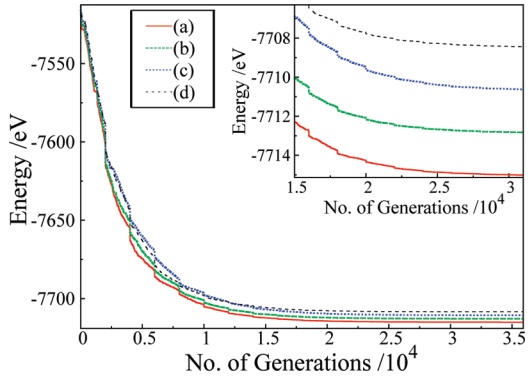


Figure 10. Energy evolution profiles of 150-ring PT chains for a (a) neutral PT chain, (b) two-bipolaron-doped PT chain, (c) four-bipolaron-doped PT chain, and (d) six-bipolaron-doped PT chain. Inset: profile in the region between 15000 and 30000 generations.

quininoid rings in the center containing about 68% of the total charge carried by a bipolaronic defect (Figure 9a). Beyond the 4 rings there is almost a continuous transition to aromatic structure which persists over 13−14 rings. Geometry optimization by the DRMHC procedure thus predicts the expected pattern of evolution of the structures of doped PTs.

The gradual increase in the doping level causes progressive reduction in the band gap, which tends to get saturated at a low but nonzero value (Figure 9b). Incorporation of a single bipolaronic defect drastically brings down the band gap ($\Delta\varepsilon$), and the Fermi level (ε_F) also moves into the valency band. The reduction in the band gap is not so spectacular as the doping level increases further. The lowest value predicted is 0.57 eV for the 12-bipolaron-doped PT chain of 100 rings. It would be of interest to find the optimized structures of bipolaron-doped PTs of even larger chain lengths to test whether our "λ-optimized DRMHC" performs well in this situation. That the geometry optimization in doped PTs proceeds smoothly is shown in representative cases in parts a−d of Figure 10 for 150-ring PT chains with zero, two, four, and six bipolarons, respectively. In each case we have taken the geometry of the neutral PT chain of 150 rings as one starting point which is far from the optimized geometry searched out by our algorithm.

4. Conclusions

λ-optimized directed random mutation hill climbing can be a viable strategy for locating minimum energy structures of undoped as well as doped polythiophenes and their analogues. The algorithm is generizable and can be used to handle the global geometry optimization problem in large molecules with complex potential energy surfaces.

Acknowledgment. K.S. thanks the CSIR, Government of India, New Delhi, for the award of a Junior Research Fellowship. We thank Dr. Pinaki Chowdhury (Calcutta University) for his help.

References

(1) Dykstra, C. E. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: New York, 2005; pp 1011−1045.

(2) Schlegel, H. B. In *Ab Initio Methods in Quantum Chemistry*; Lawley, K. P., Ed.; Advances in Chemical Physics, Vol. LXV11; John Wiley & Sons Ltd.: Chichester, U.K., 1987; pp 1−566.

(3) Cramer, C. J. *Essentials of Computational Chemistry Theories and Models*, 2nd ed.; John Wiley & Sons Ltd.: Chichester, U.K., 2004; pp 487−516.

(4) Kobayashi, M.; Chen, J.; Chung, T. C.; Heeger, A. J.; Wudl, F. Synthesis and properties of chemically coupled poly(thiophene). *Synth. Met.* **1984**, *9*, 77–86.

(5) Logdlund, M.; Lazzaroni, R.; Stafström, S.; Salaneck, W. R.; Bredas, J. L. Direct observation of charge-induced π-electronic structural changes in a conjugated polymer. *Phys. Rev. Lett.* **1989**, *63*, 1841–1844.

(6) Stafström, S.; Bredas, J. L. Evolution of the electronic structure of polyacetylene and polythiophene as a function of doping level and lattice conformation. *Phys. Rev. B* **1988**, *38*, 4180–4191.

(7) Giri, D.; Kundu, K. Theoretical study of the evolution of electronic band structure of polythiophene due to bipolaron doping. *Phys. Rev. B* **1996**, *53*, 4340–4350.

(8) Mazziotti, D. A. Anti-Hermitian contracted Schrödinger equation: Direct determination of the two-electron reduced density matrices of many-electron molecules. *Phys. Rev. Lett.* **2006**, *97*, 143002.

(9) Mazziotti, D. A. Anti-Hermitian part of the contracted Schrödinger equation for the direct calculation of two-electron reduced density matrices. *Phys. Rev. A* **2007**, *75*, 022505.

(10) Gidofalvi, G.; Mazziotti, D. A. Direct calculation of excited-state electronic energies and two-electron reduced density matrices from the anti-Hermitian contracted Schrödinger equation. *Phys. Rev. A* **2009**, *80*, 022507.

(11) Lavadra, F. C.; dos Santos, M. C.; Galvao, D. S.; Lak, B. Insulator-to-metal transition in polythiophene. *Phys. Rev. B* **1994**, *49*, 979–983.

(12) Su, W. P.; Schrieffer, J. R.; Heeger, A. J. Solitons in polyacetylene. *Phys. Rev. Lett.* **1979**, *42*, 1698–1701.

(13) Su, W. P.; Schrieffer, J. R.; Heeger, A. J. Soliton excitations in polyacetylene. *Phys. Rev. B* **1980**, *22*, 2099–2111.

(14) Mitchel, M.; Forrest, S.; Holland, J. H. In *When Will a Genetic Algorithm Outperform Hill Climbing?*; Cowen, J. D., Tesauro, G., Alspector, J., Eds.; Advances in Neural Information Processing Systems, Vol. 6; Morgan Kaufmann: San Mateo, CA, 1994.

(15) Mitchell, M. *An Introduction to Genetic Algorithms*, 1st ed.; MIT Press: Cambridge, MA, 1996; pp 128−132.

(16) Chaudhury, P.; Bhattacharyya, S. P. Numerical solutions of the Schrödinger equation directly or perturbatively by a genetic algorithm: Test cases. *Chem. Phys. Lett.* **1998**, *296*, 51–60.

(17) Sharma, R.; Bhattacharyya, S. P. Direct search for wave operator by a genetic algorithm (GA): Route to few eigenvalues of a Hamiltonian. *Proceedings of the IEEE Congress on Evolutionary Computation, Singapore*; IEEE Press: Piscataway, NJ, 2007; pp 3812−3817.

(18) Nandy, S.; Chaudhury, P.; Sharma, R.; Bhattacharyya, S. P. A density-matrix-based simulated annealing (SA) technique for locating minimum energy structures on the neutral polythiophene potential energy surface. *J. Theor. Comput. Chem* **2008**, *7*, 977–987.

(19) Schaffer, J. D.; Caruana, R. A.; Eshelman, L. J.; Das, R. A study of control parameters affecting online performance of genetic algorithms for function optimization. *Proc. 3rd Int. Conf. Genet. Algorithms* **1989**, 51–60.

(20) Hesser, J.; Manner R. Towards an optimal mutation probability in genetic algorithms. In *Parallel Problem Solving from Nature*; Schwefel, H.-P., Manner, R., Eds.; Lecture Notes in Computer Science, Vol. 496; Springer: Berlin, 1991; pp 23−32.

(21) Bredas, J. L.; Street, G. B. Polarons, Bipolarons, and Solitons in Conducting Polymers. *Acc. Chem. Res.* **1985**, *18*, 309–315.

CT900540D

JCTC Journal of Chemical Theory and Computation

An Assessment of Density Functional Methods for Potential Energy Curves of Nonbonded Interactions: The XYG3 and B97-D Approximations

Álvaro Vázquez-Mayagoitia,[†] C. David Sherrill,*[,‡] Edoardo Aprà,[§] and Bobby G. Sumpter[§]

Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry and College of Computing, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, Chemistry Department, University of Tennessee, 1416 Circle Drive, 552 Dabney-Buehler Hall, Knoxville, Tennessee 37996-1600, and Computer Science and Mathematics Division and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Bethel Valley Road, P.O. Box 2008, Building 6012, Oak Ridge, Tennessee 37831-6367

Received October 17, 2009

Abstract: A recently proposed double-hybrid functional called XYG3 and a semilocal GGA functional (B97-D) with a semiempirical correction for van der Waals interactions have been applied to study the potential energy curves along the dissociation coordinates of weakly bound pairs of molecules governed by London dispersion and induced dipole forces. Molecules treated in this work were the parallel sandwich, T-shaped, and parallel-displaced benzene dimer, $(C_6H_6)_2$; hydrogen sulfide and benzene, $H_2S \cdot C_6H_6$; methane and benzene, $CH_4 \cdot C_6H_6$; the methane dimer, $(CH_4)_2$; and the pyridine dimer, $(C_5H_5N)_2$. We compared the potential energy curves of these functionals with previously published benchmarks at the coupled cluster singles, doubles, and perturbative triplets [CCSD(T)] complete-basis-set limit. Both functionals, XYG3 and B97-D, exhibited very good performance, reproducing accurate energies for equilibrium distances and a smooth behavior along the dissociation coordinate. Overall, we found an agreement within a few tenths of one kcal mol^{-1} with the CCSD(T) results across the potential energy curves.

I. Introduction

Nonbonded interactions strongly affect protein folding, DNA structure, supramolecular assembly, and drug docking.[1−4] Indeed, any attempt to perform simulations on condensed matter systems is severely handicapped without an adequate description of such interactions. While nonbonded interactions are approximately described by empirical force field methods, these methods are not always accurate for some types of nonbonded interactions, including $\pi-\pi$ stacking.[5] Moreover, even ab initio methods can exhibit significant errors for nonbonded interactions[6−8] unless very large basis sets are used in conjunction with highly correlated methods such as coupled-cluster with perturbative triple excitations, CCSD(T).[9] Unfortunately, most popular density functional approximations completely fail to describe long-range London dispersion interactions.[7,10,11]

Numerous approaches have been advanced to overcome these failures. Of the proposed solutions, the simplest conceptually is to add a damped, empirical dispersion term, yielding a method usually designated as DFT-D.[12−16] This approach seems to give reliable results for nonbonded interactions in a variety of geometries.[17,18] However, while the approach is simple and cost-effective, the use of empirical terms is theoretically unappealing, and the density is not coupled to the dispersion interaction. Other workers have tried reparametrization or extensions of existing types of functionals[19−23] with some success, but ultimately the

* Corresponding author e-mail: sherrill@gatech.edu.
[†] University of Tennessee.
[‡] Georgia Institute of Technology.
[§] Oak Ridge National Laboratory.

10.1021/ct900551z
Published on Web 02/01/2010

physics of long-range dispersion is not captured by local or semilocal functionals.[10] More rigorous approaches include the addition of nonlocal terms to the functional,[24−26] and the exchange dipole moment method of Becke and Johnson.[27−29]

Recently, Zhang, Xu, and Goddard have introduced a "double hybrid" density functional, which mixes in Hartree−Fock exchange as well as contributions from unoccupied orbitals via second-order perturbation theory.[30] Their empirical XYG3 functional uses a scheme of three parameters similar to the B3LYP functional.[31,32] Here, we examine the performance of this double hybrid XYG3 functional for potential energy curves of prototypical nonbonded interactions, and we compare the performance of the XYG3 functional to that of the dispersion-corrected B97-D functional introduced by Grimme.[15] Assessing the quality of these approximations across a range of intermolecular separations is important because, in larger complexes, noncovalent interactions between chemical groups will occur in a wide variety of geometries, and the number of long-range contacts will grow with respect to system size.

II. Theoretical and Computational Details

Density Ffnctional theory (DFT)[33,34] proposes to solve electronic structure problems using as a fundamental variable the electron charge density, $\rho(\vec{r})$; formally, it is based on the Hohenberg and Kohn theorems.[35] In practice, DFT is applied using the Kohn−Sham method (KS), using a mean-field approach.[36] The KS method represents the density as a linear combination of the inner products of spin−orbital functions $\rho(\vec{r}) = \sum_i^{N_{ele}} |\psi_i^{KS}(\vec{r})|^2$, and the energy as a functional of $\rho(\vec{r})$ as

$$E^{DFT}[\rho] = T_0[\rho] + J[\rho] + E_{xc}[\rho] + \int d\vec{r}\, \rho(\vec{r}) v_{ext}(\vec{r}) + V_{NN} \quad (1)$$

where the first and second terms are the kinetic energy of independent particles, $T_0[\rho]$, and the Coulomb interaction energy, $J[\rho] = 1/2 \iint d\vec{r}'\, d\vec{r} \rho(\vec{r}')\rho(\vec{r})/|\vec{r}' - \vec{r}|$. The term $v_{ext}(\vec{r})$ is the external potential generated by the nuclei and felt by the electrons, and V_{NN} is the nuclear repulsion energy for a fixed nuclear configuration. In eq 1, the contribution $E_{xc}[\rho]$ is the exchange-correlation energy, which includes the electron exchange interaction as well as the many-body contribution to the kinetic and electron−electron repulsion potentials ($V_{ee}[\rho]$) that are not included in $T_0[\rho]$ or $J[\rho]$, that is, $E_{xc}[\rho] = V_{ee}[\rho] + T[\rho] - J[\rho] - T_0[\rho]$. The explicit expression of E_{xc} remains unknown, but there are many approaches that have shown satisfactory results. Such approaches have been grouped according to their treatment of the density into "generations" or "ladder's rungs".[37] The most used are based on the Local Density Approximation (LDA) or Generalized Gradient Approximation (GGA).[38] Although some functionals have shown impressive results, those are not totally transferable for every problem and especially fail for the description of long-range interactions and excited states. The origins of these difficulties are attributed to the incorrect cancellation of electron self-interaction,[39] and incorrect treatment of dynamic correlation, among others. There are many strategies to avoid these problems, some of which involve the inclusion of explicit terms from wave function theories (hybrid functionals),[31,40] treatments with optimized effective potentials,[41−43] an adjustment to the asymptotic correction exchange correlation potentials,[44] and the addition of empirical energy terms.

A. Double Hybrid Functionals. The (double) hybrid functionals emerge from the adiabatic connection formalism[31,45] where the exchange correlation functional is obtained solving the follow equation:

$$E_{xc} = \int_0^1 d\lambda\, U_{xc}^{\lambda} \quad (2)$$

The integrand $U_{xc}^{\lambda}[\rho]$ is an exchange-correlation potential energy, keeping fixed the $v_{ext}(\vec{r})$ and $\rho(\vec{r})$ of the physical system, and depending on a dimensionless interelectronic coupling-strength constant, λ, that switches smoothly between a model of independent particles and one of interacting particles around an interval $0 \le \lambda \le 1$, using the KS orbitals as the reference system.

The integrand of eq 2 can be expressed as $U_{xc}^{\lambda} = V_{ee}^{\lambda}[\rho] - J[\rho] + T^{\lambda}[\rho] - T_0[\rho]$; thus this potential depends on λ by virtue of the Hellmann−Feyman theorem. Independently of the form of U_{xc}^{λ}, it is possible to establish certain boundary conditions. The lower integral limit represents the exact exchange, $U_{xc}^{\lambda=0} = V_{ee}^{\lambda=0}[\rho] - J[\rho] + T^{\lambda=0}[\rho] - T_0[\rho] = E_x$, and its upper limit $U_{xc}^{\lambda=1} = V_{ee}^{\lambda=1}[\rho] - J[\rho] + T^{\lambda=1}[\rho] - T_0[\rho] = E_{xc}$, which could be one of the LDA or GGA exchange-correlation energies represented by $U_{xc}^{\lambda=1}[\rho] \approx E_{xc}^{DFT}[\rho]$.

Alternatively, invoking the perturbation scheme proposed by Görling and Levy (GLPT),[46,47] the correlation energy is expanded in a power series around $\lambda = 0$. The first order corresponds to E_x and the second as E_c^{GLPT}; in this formulation, the gradient at $\lambda = 0$ is equal to $(\partial U_{xc}^{\lambda}/\partial\lambda)_{\lambda=0} = 2E_c^{GLPT}$. The E_c^{GLPT} correlation functional explicitly includes single and double orbital transitions, written in Mulliken notation for the electron repulsion integrals as

$$E_c^{GLPT} = -\sum_{ai} \frac{\left|\langle a|\hat{v}_x|i\rangle - \sum_j (aj|ij)\right|^2}{\varepsilon_a - \varepsilon_i} - \frac{1}{4}\sum_{ijab} \frac{[(ia|jb) - (ib|ja)]^2}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j} \quad (3)$$

where i, j are defined as occupied and a, b as unoccupied KS orbitals, and the exchange potential as $\hat{v}_x = (\delta E_x^{DFT}[\rho]/\delta\rho)$.

Using a first-order interpolator pathway,[48] the integrand in eq 2 can be written as

$$U_{xc}^{\lambda} = a + b\lambda \quad (4)$$

To conciliate the boundary conditions and the correlation formulated by GLPT, an ansatz is used with the parametrical combination of the energy E_c^{GLPT} and E_{xc}^{DFT} within the LDA or GGA exchange-correlation functionals, so the constant b in eq 4 is split into two parameters for the correlation, where the final equation is given by

$$E_{xc} = a_x E_x^{HF} + (1 - a_x)E_x^{DFT} + b_1 E_c^{PT2} + b_2 E_c^{DFT} \quad (5)$$

The term E_c^{PT2} is the correlation energy established from the Møller−Plesset second-order perturbation theory, related to second term in eq 3. The single excitations of E_c^{GLPT} are neglected or simply absorbed by E_c^{DFT}.

The functionals like eq 5 are considered as a last generation and are known as double-hybrid functionals, mainly due to their direct dependence on the occupied and virtual KS orbitals, which offer a way to include the dynamic correlation explicitly, and were originally proposed by Grimme within his B2PLYP functional.[49] The potential generated by the E_c^{GLPT} term acts as a multiplicative operator and can be inserted into the mean-field equation solution via the exchange correlation potential ($\hat{v}_{xc}^{GLPT} = \delta E_{xc}^{GLPT}[\rho]/\delta\rho$) using an optimized effective potential scheme (OEP).[50−53] However, in practice, the orbitals that are used to evaluate the E_c^{PT2} typically come from a mean-field procedure that minimizes the energy for the rest of terms in eq 5.

Zhang, Xu, and Goddard have proposed[30] a set of empirical parameters for a double-hybrid functional called XYG3.

$$E_{xc}^{XYG3} = a_x E_x^{HF} + (1 - a_x)E_x^{Slater} + a_0 \Delta E_x^{B88} + a_c E_c^{PT2} + (1 - a_c)E_c^{LYP} \quad (6)$$

In this case, they adopt the three empirical terms $a_x = 0.8033$, $a_0 = 0.2107$, and $a_c = 0.3211$ in the same spirit as the B3LYP functional, with values that best fit thermodynamical data for the G3/99 set.[54] The functionals E_x^{B88} and E_c^{LYP} denote an exchange functional by Becke[55] and a correlation by Lee, Yang, and Parr, respectively.[56] The main difference with respect to other reparametrizations such as B2K-PLYP,[57] B2GP-PLYP,[58] or B2-P3LYP,[59] which also can produce thermodynamical data and reactions barriers quite similar to those from high-level methods, is that the XYG3 functional adjusts the gradient correction using the parameter a_0 in eq 6, adding the E_x^{Slater} exchange. The XYG3 functional has been tested using the electron density and orbital functions from B3LYP, and this approximation is already capable of accurately reproducing heats of formation, energy barriers, and noncovalent interactions with very good results, and as shown below it can reproduce potential energy surfaces of weak interactions. Although we focus on the XYG3 double hybrid functional in this work, we also discuss limited results for the original B2PLYP double hybrid[49] as well as its empirically corrected variant, B2PLYP-D.[60]

B. Semiempirical Dispersion Contribution. Standard density functionals are local or at best semilocal, and hence they neglect long-range electron correlations, which give rise to attractive London dispersion forces. It has been proposed to simply add an empirical term to account for the missing dispersion energies, that is,

$$E_{DFT\text{-}D} = E^{DFT} + E_{dispersion} \quad (7)$$

From observations, is well-known that the dispersion energy contributes asymptotically to the potential energy in long-range interactions as $U_{disp} \approx -R^{-6}$.[61] Thus, modeling of the dispersion energy as the interaction between pairs of atoms was proposed:[12−16]

$$E_{dispersion} = -s_6 \sum_{i=1}^{N_{at}} \sum_{j>i}^{N_{at}} f(R_{ij}) \frac{C_6^{ij}}{R_{ij}^6} \quad (8)$$

Here, the function $f(R_{ij})$ acts as a damping function, with a gradual transition around a scaled distance R_{ij}^0, which is the sum of individual atomic van der Waals radii, $R_i^0 + R_j^0$. This function is modeled as a Fermi−Dirac-like distribution, $f(R_{ij}) = \{1 + \exp[-\alpha(R_{ij}/R_{ij}^0 - 1])\}^{-1}$, under the control of a global α parameter. The scalar s_6 value in eq 8 weights the whole contribution and is adjusted parametrically for each E_{xc}^{DFT} functional. Furthermore, C_6^{ij} can be determined by an average between the C_6 of i, j atoms, frequently using a geometric mean, $C_6^{ij} = \sqrt{(C_6^i C_6^j)}$.

In this work, we use the E_{xc} proposed by Grimme[15] known as B97-D. This functional is based on a previous one proposed by Becke.[62] Essentially B97-D is a reparametrization to coefficients of an expansion series with a gradient correction factor inside a E_{xc}^{GGA}. The coefficients are optimized by a least-squares fitting procedure, including the term in eq 8, to best reproduce heats of formation from the G2/97 set[63,64] and other properties such as ionization potentials, atomization energies, etc. Additionally, the B97-D functional also attempts to improve the short-range description and avoid the double-counting of electron correlation at medium range distances when the dispersion correction is present. Finally, a remarkable point is that B97-D does not have the E_x^{HF} energy as B97 does, which allows a significant reduction in computational effort, especially when using auxiliary fitted basis functions[65] to evaluate the two-electron integrals.

C. Computational Methods. Computations employed the triple-ζ basis sets used during the development of the functionals considered, TZV2P for B97-D and 6-311+G(3df,2p) for XYG3. We anticipate that the functionals will work best when paired with these basis sets, which were used during their parametrization. Limited tests indicate that the B97-D results are fairly insensitive to basis set, and we found very similar results when using the TZV2P basis or Dunning's aug-cc-pVTZ basis.[66] Earlier work indicates that the typical DFT-D methods are also rather insensitive to basis set superposition error,[15] making counterpoise correction[67] unnecessary. Our results here are not counterpoise corrected unless otherwise noted. The double-hybrid functionals are somewhat more sensitive to basis set and exhibit larger basis set superposition errors because of the perturbation theory term, as we discuss in more detail below.

Both functionals were implemented into the quantum chemistry code NWChem.[68] For the XYG3 functional, we used for the second-order perturbation part the KS orbitals that minimize the energy of the B3LYP functional. For comparison purposes, some results are also obtained using counterpoise-corrected second-order Møller−Plesset perturbation theory (MP2) with an aug-cc-pVDZ basis.

Potential energy curves are evaluated for several prototype dimers for which high-accuracy estimates quantum mechanical benchmarks are available. The benchmark data were obtained by extrapolating MP2 to the complete-basis-set

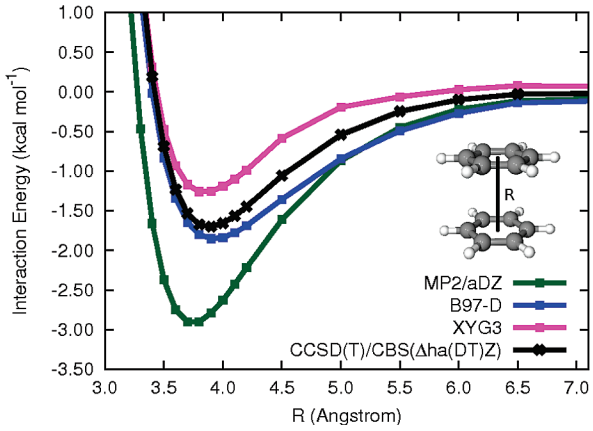


Figure 1. Potential energy curves for the sandwich benzene dimer. CCSD(T)/CBS results from ref 18.

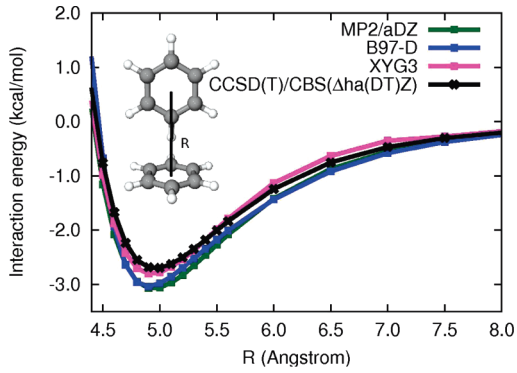


Figure 2. Potential energy curves for the T-shaped benzene dimer. CCSD(T)/CBS results from ref 18.

(CBS) limit, and adding a higher-order electron correlation correction evaluated as the difference between CCSD(T) and MP2 using smaller basis sets. We will denote the basis set(s) used for this "ΔCCSD(T)" correction in parentheses, so that, for example, CCSD(T)/CBS(Δha(DT)Z) denotes an estimate obtained from MP2/CBS binding energies, with a ΔCCSD(T) correction evaluated as the difference between CCSD(T) and MP2, both extrapolated to the CBS limit using a heavy-aug-cc-pVDZ/heavy-aug-cc-pVTZ two-point extrapolation[69] (and where "heavy-aug" indicates only non-hydrogen atoms are augmented by diffuse functions). Recent work indicates that such approximation schemes are very effective at approaching the true CCSD(T)/CBS limit, especially when it is possible to use basis sets larger than augmented double-ζ for the ΔCCSD(T) correction.[18]

Here, we use previously published CCSD(T)/CBS estimates as benchmarks for the following systems: the sandwich, T-shaped, and parallel-displaced configurations of the benzene dimer,[18] methane dimer,[70] methane-benzene,[18] H_2S−benzene,[18] the antiparallel sandwich pyridine dimer,[71] and a T-shaped pyridine dimer.[71]

III. Results and Discussion

Potential energy curves (relative to infinitely separated monomers) are presented in Figures 1−10 for B97-D, XYG3, and estimated CCSD(T)/CBS. The benzene dimer figures also include (counterpoise corrected) MP2/aug-cc-pVDZ results for comparison. Table 1 presents the benchmark

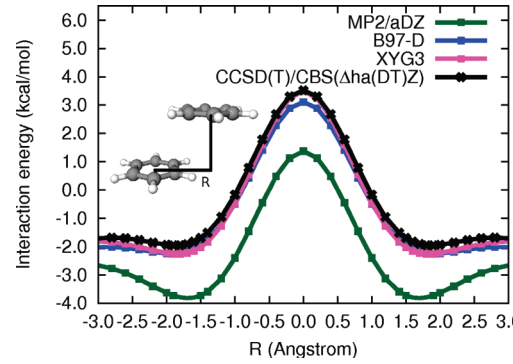


Figure 3. Potential energy curves for the parallel-displaced benzene dimer, with a vertical separation of 3.2 Å. CCSD(T)/CBS results from ref 18.

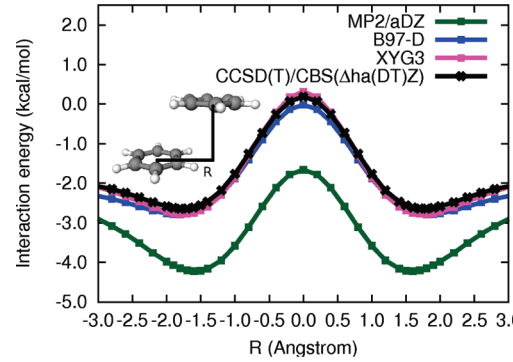


Figure 4. Potential energy curves for the parallel-displaced benzene dimer, with a vertical separation of 3.4 Å. CCSD(T)/CBS results from ref 18.

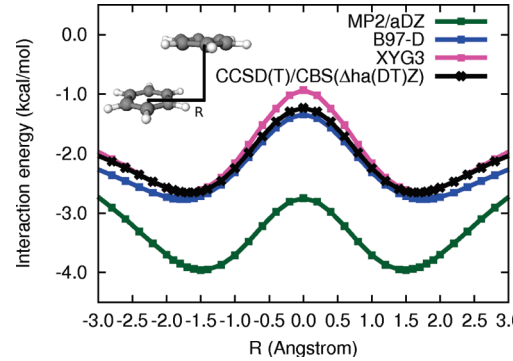


Figure 5. Potential energy curves for the parallel-displaced benzene dimer, with a vertical separation of 3.6 Å. CCSD(T)/CBS results from ref 18.

CCSD(T)/CBS equilibrium intermolecular distances and interaction energies for each system, and the errors in these quantities for each of the approximate methods considered. In general, both the B97-D and the XYG3 methods yield quantitatively reliable potential energy curves. MP2/aug-cc-pVDZ, by contrast, yields qualitatively correct but quantitatively poor curves, which exhibit significant overbinding for the sandwich and parallel-displaced configurations of the benzene dimer. Larger basis sets lead to even greater overbinding by MP2. Because of these fairly large errors, MP2 results are not displayed for the remaining test cases so that the range of the graphs makes the errors for B97-D and XYG3 easier to see.

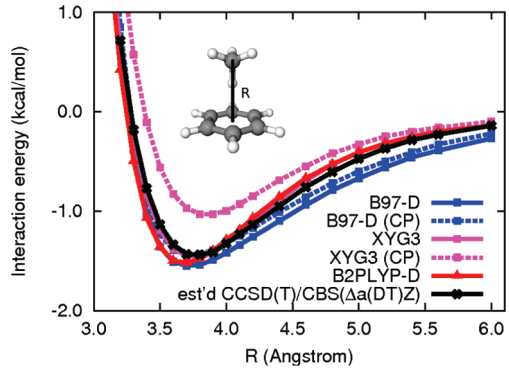


Figure 6. Potential energy curves for the methane−benzene complex. CCSD(T)/CBS results from ref 18. Curves labeled (CP) include counterpoise correction. The XYG3 curve is essentially coincident with the CCSD(T) curve to the left of equilibrium, and coincident with the B2PLYP-D curve to the right of equilibrium.

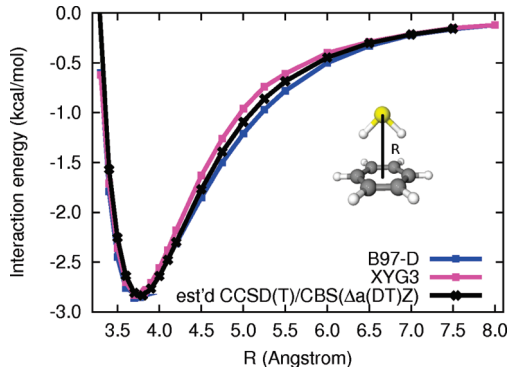


Figure 7. Potential energy curves for the H_2S−benzene complex. CCSD(T)/CBS results from ref 70.

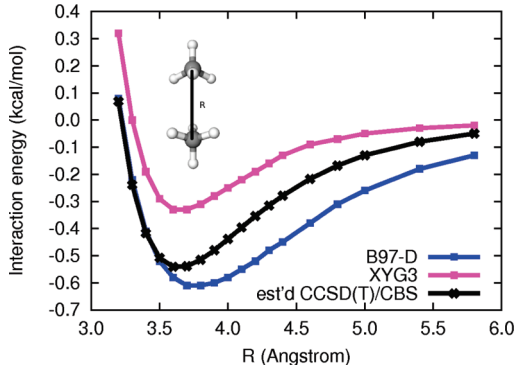


Figure 8. Potential energy curves for the methane dimer. CCSD(T)/CBS results from ref 70.

In the case of the benzene−methane complex (Figure 6), we have examined the effect of counterpoise correction and have compared the XYG3 double hybrid to the B2PLYP-D empirically corrected double hybrid[60] with the TZV2P basis set. We also evaluated the original B2PLYP double hybrid for this case (results not pictured), but B2PLYP-D performs significantly better (the B2PLYP errors are around 0.8 kcal mol^{-1} near equilibrium, as compared to around 0.1 kcal mol^{-1} for B2PLYP-D). For this case, B2PLYP-D and XYG3 are nearly identical. Considering the effect of basis set superposition error, B97-D is hardly affected by counterpoise correction, but the counterpoise-corrected XYG3 curve is significantly underbound and much less accurate than the uncorrected XYG3. This suggests that XYG3 will perform best when used without counterpoise correction and with the 6-311+G(3df,2p) basis set used during its parametrization.

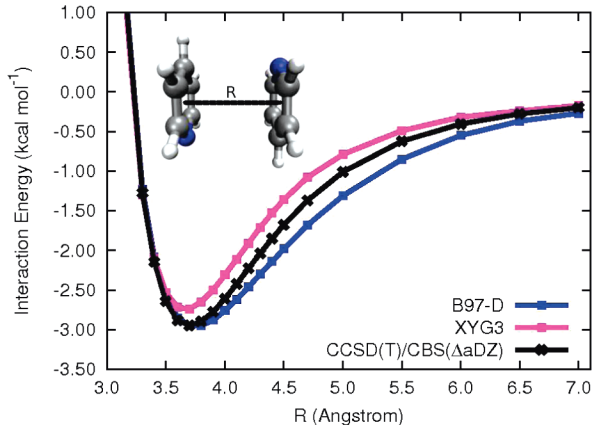


Figure 9. Potential energy curves for the antiparallel sandwich pyridine dimer. CCSD(T)/CBS results from ref 71.

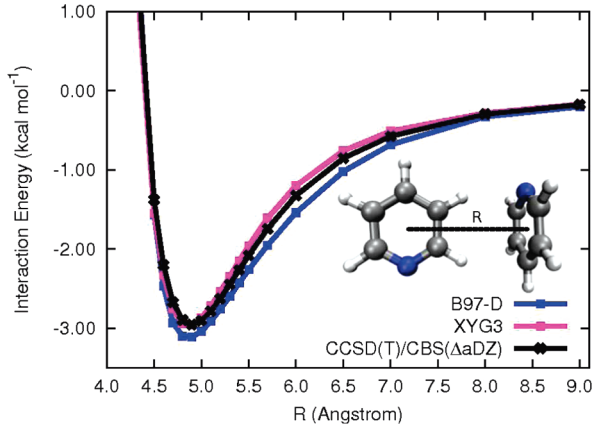


Figure 10. Potential energy curves for a T-shaped pyridine dimer. CCSD(T)/CBS results from ref 71.

Table 1. CCSD(T) Equilibrium Intermolecular Distances (angstroms) and Interaction Energies (kcal mol^{-1}), and Errors for XYG3 and B97-D

	CCSD(T)		XYG3		B97-D	
dimer	R	E_{int}	ΔR	ΔE_{int}	ΔR	ΔE_{int}
benzene dimer sandwich	3.9	−1.701	−0.1	0.443	0.0	−0.151
benzene dimer T-shaped	5.0	−2.698	−0.1	−0.106	−0.1	−0.340
benzene dimer PD (3.2 Å)	1.9	−1.957	0.1	−0.301	0.0	−0.289
benzene dimer PD (3.4 Å)	1.8	−2.643	0.0	−0.147	−0.1	−0.140
benzene dimer PD (3.6 Å)	1.7	−2.654	0.0	−0.002	−0.1	−0.115
CH_4−benzene	3.8	−1.439	−0.1	−0.011	−0.1	−0.111
H_2S−benzene	3.8	−2.834	−0.1	0.004	−0.1	−0.026
CH_4−CH_4	3.6	−0.541	0.0	0.211	0.1	−0.069
pyridine dimer sandwich	3.7	−2.948	0.0	0.208	0.0	−0.002
pyridine dimer T-shaped	4.9	−2.954	0.0	−0.006	0.0	−0.156
mean deviation			0.0	0.029	0.0	−0.140
mean absolute deviation			0.1	0.149	0.1	0.140

For XYG3, the most noticeable errors are for the sandwich benzene dimer and for the methane dimer. In the sandwich benzene dimer, the XYG3 potential well is too shallow by about 0.5 kcal mol^{-1}, and it remains underbound by several tenths of one kcal mol^{-1} at larger distances also. The methane dimer, likewise, is underbound across the potential curve, with errors around 0.2 kcal mol^{-1} near equilibrium; this error decreases slowly as the intermolecular distance increases. These errors are quite small, but they are perhaps significant when compared to the total methane dimer CCSD(T)/CBS interaction energy of −0.54 kcal mol^{-1} at equilibrium. XYG3 shows slight underbinding in the parallel benzene dimer at small horizontal displacements (near the sandwich configuration, amouting to an error of about 0.3 kcal mol^{-1} at a vertical separation of 3.6 Å), and for the antiparallel pyridine sandwich, there is also a slight underestimation of the intermolecular interation (by about 0.2 kcal mol^{-1} near equilibrium). Although somewhat difficult to see from Figure 3, XYG3 is slightly overbound for the parallel-displaced benzene dimer, with errors of around −0.3 kcal mol^{-1} for intermediate horizontal displacements. For the other test cases, the XYG3 potential curves lie very close to the benchmark CCSD(T)/CBS curves. Overall, XYG3 performs very well, with errors of a few tenths of a kcal mol^{-1} or less across the potential curves.

The overall performance of B97-D is quite similar to that of XYG3; it also yields binding energies that are within a few tenths of one kcal mol^{-1} of CCSD(T)/CBS values across the potential curves, but it tends to slightly overestimate binding. B97-D is somewhat more reliable than XYG3 for estimating interaction energies near equilibrium geometries of several of the systems considered here (e.g., sandwich benzene dimer, methane dimer, and the antiparallel sandwich pyridine dimer), but it does not perform as well as XYG3 at the equilibrium geometry of the T-shaped benzene dimer or the T-shaped pyridine dimer. Overall, the mean absolute errors in equilibrium interaction energies are very similar for XYG3 and B97-D (0.15 and 0.14 kcal mol^{-1}, respectively; see Table 1). In a majority of the test cases, B97-D slightly overestimates binding at intermediate to large intermolecular distances.[72] A typical error would be in the range of 0−0.2 kcal mol^{-1}, although it can be larger (e.g., around 0.3 kcal mol^{-1} for the sandwich benzene dimer at an intermolecular separation of 5.0 Å).

Table 2 presents the statistics for the errors across the potential curves considered. Although the more noticeable errors for XYG3 in the figures correspond to a slight underbinding, the mean error across all points considered is actually very close to zero (about 0.01 kcal mol^{-1}). The mean absolute deviation across all points is 0.14, and the maximum error is 0.47 kcal mol^{-1}. As mentioned above, B97-D has a slight tendency to overbind dimers, and this is reflected in the negative mean error in interaction energies of −0.16 kcal mol^{-1}. The mean absolute deviation of 0.18 and the maximum error of 0.68 kcal mol^{-1} are both slightly larger than those observed for XYG3 for the points considered. Overall, then, XYG3 and B97-D appear to perform similarly for these protoypes of nonbonded interactions, with the errors of XYG3 being slightly smaller.

Table 2. Statistics for the Errors of Approximate Methods across the Potential Energy Curves Considered[a]

		XYG3			B97-D		
curve	N_{pts}	MD	MAD	MAX	MD	MAD	MAX
benzene dimer sandwich	17	0.27	0.27	0.47	−0.20	0.20	−0.41
benzene dimer T-shaped	18	−0.05	0.10	−0.24	−0.15	0.22	0.57
benzene dimer PD (3.2 Å)	37	−0.24	0.24	−0.33	−0.32	0.32	−0.41
benzene dimer PD (3.4 Å)	37	−0.07	0.10	−0.15	−0.17	0.17	−0.24
benzene dimer PD (3.6 Å)	37	0.07	0.07	0.30	−0.13	0.13	−0.23
CH_4−benzene	19	0.01	0.04	−0.08	−0.15	0.15	−0.25
H_2S−benzene	19	0.03	0.08	−0.18	−0.06	0.07	−0.22
CH_4−CH_4	18	0.17	0.17	0.25	−0.10	0.10	−0.17
pyridine dimer sandwich	20	0.18	0.20	0.33	−0.12	0.15	−0.31
pyridine dimer T-shaped	19	−0.00	0.10	−0.33	−0.10	0.19	0.68
total	241	0.01	0.14	0.47	−0.16	0.18	0.68

[a] N_{pts} denotes the number of points along the curve. Mean deviation (MD), mean absolute deviation (MAD), and maximum errors (MAX) are given in kcal mol^{-1}.

IV. Conclusions

The recent availability of benchmark-quality potential energy curves for various weakly bound dimers makes it possible to assess the quality of new theoretical methods for describing nonbonded interactions across a range of geometries. Here, we compared the double-hybrid density functional approximation XYG3 and the empirically dispersion-corrected B97-D approach. Both methods performed very well for the potential energy curves considered, with errors no more than a few tenths of one kcal mol^{-1} along the curves. B97-D with its recommended TZV2P basis tends to slightly overbind the dimers considered, while XYG3 with its recommended 6-311+G(3df,2p) basis overbinds about as often as it underbinds. For the benzene−methane complex, we also examined the empirically corrected double-hybrid functional B2PLYP-D,[60] which performed nearly identically to XYG3. For this dimer, we also found that B97-D/TZV2P is rather insensitive to counterpoise correction, whereas counterpoise-corrected XYG3/6-311+G(3df,2p) results were significantly less accurate than the uncorrected values.

The B97-D and XYG3 methods exhibited mean absolute deviations of 0.18 and 0.14 kcal mol^{-1}, respectively, across all geometries considered. Such small errors are difficult to achieve even when using high-level electronic structure methods for these systems,[18] and thus either approach appears to be suitable for typical applications to weakly interacting systems. It should be noted that double-hybrid functionals have a significantly greater computational cost as compared to pure DFT functionals, although in principle this could be alleviated to some extent by using an auxiliary basis and local correlation approaches.

Acknowledgment. We thank Michael S. Marshall for assistance with the figures. This work was supported by the National Science Foundation (Grant No. CHE-0715268) and by the Donors of the American Chemical Society Petroleum Research Fund (Grant No. 44262-AC6). The Center for

Computational Molecular Science and Technology is funded through an NSF CRIF award (CHE 04-43564) and by Georgia Tech. B.G.S. acknowledges support by the Center for Nanophase Materials Sciences (CNMS), sponsored by the Division of Scientific User Facilities, U.S. Department of Energy. E.A. and A.V.M. acknowledge support from the Department of Energy, Offices of Basic Energy Science and Advanced Scientific Computing Research as part of the SciDAC program. This research used resources supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Supporting Information Available: Cartesian coordinates for equilibrium geometries, and potential energy curves of the complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Steed, J. W.; Atwood, J. L. *Supramolecular Chemistry: A Concise Introduction*; Wiley: New York, 2000.

(2) Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives*; VCH: New York, 1995.

(3) Diederich, F., Künzer, H., Eds. *Recent Trends in Molecular Recognition*; Springer: New York, 1998.

(4) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210–1250.

(5) Sherrill, C. D.; Sumpter, B. G.; Sinnokrot, M. O.; Marshall, M. S.; Hohenstein, E. G.; Walker, R. C.; Gould, I. R. *J. Comput. Chem.* **2009**, *30*, 2187–2193.

(6) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790–18794.

(7) Tsuzuki, S.; Lüthi, H. P. *J. Chem. Phys.* **2001**, *114*, 3949.

(8) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656–10668.

(9) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(10) Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.

(11) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334–338.

(12) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

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

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

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

(16) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.

(17) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. *J. Chem. Theory Comput.* **2008**, *4*, 1996–2000.

(18) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.

(19) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *101*, 2673–2677.

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

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

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

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

(24) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004–153007.

(25) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. B* **2005**, *71*, 195119.

(26) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401–246404.

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

(28) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.

(29) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104.

(30) Zhang, Y.; Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4963–4968.

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

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

(33) Parr, R.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(34) Kohanoff, J. *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods*; Cambridge University Press: New York, 2006.

(35) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871.

(36) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(37) Perdew, J. P.; Schmidt, K. *Jacob's Ladder of Density Functional Approximations for the Exchange-Correlation Energy*; AIP: New York, 2001; Vol. 577.

(38) Perdew, J.; Kurth, S. *A Primer in Density Functional Theory*; Springer Verlag: Berlin Heidelberg, 2003.

(39) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.

(40) Csonka, G.; Johnson, B. *Theor. Chem. Acc.* **1998**, *99*, 158–165.

(41) Krieger, J. B.; Li, Y.; Iafrate, G. J. *Phys. Rev. A* **1992**, *46*, 5453–5458.

(42) Krieger, J.; Li, Y.; Iafrate, G. J. *Phys. Rev. A* **1992**, *45*, 101–126.

(43) Garza, J.; Nichols, J. A.; Dixon, D. A. *J. Chem. Phys.* **2000**, *112*, 7880.

(44) Levy, M.; Perdew, J. P.; Sahni, V. *Phys. Rev. A* **1984**, *30*, 2745–2748.

(45) Frydel, D.; Terilla, W. M.; Burke, K. *J. Chem. Phys.* **2000**, *112*, 5292–5297.

(46) Görling, A.; Levy, M. *Phys. Rev. B* **1993**, *47*, 13105.

(47) Teale, A. M.; Coriani, S.; Helgaker, T. *J. Chem. Phys.* **2009**, *130*, 104111.

(48) Peach, M. J. G.; Miller, A. M.; Teale, A. M.; Tozer, D. J. *J. Chem. Phys.* **2008**, *129*, 064105.

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

(50) Mori-Sánchez, P.; Wu, Q.; Yang, W. *J. Chem. Phys.* **2005**, *123*, 062204.

(51) Facco Bonetti, A.; Engel, E.; Schmid, R. N.; Dreizler, R. M. *Phys. Rev. Lett* **2001**, *86*, 2241–2244.

(52) Grabowski, I.; Hirata, S.; Ivanov, S.; Bartlett, R. J. *J. Chem. Phys.* **2002**, *116*, 4415–4425.

(53) Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3–60.

(54) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *Chem. Phys. Lett.* **1999**, *313*, 600–607.

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

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

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

(58) Karton, A.; Tarnopolsky, A.; Lamére, J.-F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.

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

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

(61) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press: New York, 1997.

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

(63) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(64) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 42–55.

(65) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

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

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

(68) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, E.; Windus, T. L.; Nichols, J. H. 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.; Voorhis, T. V.; 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, WA, 2007.

(69) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(70) Takatani, T.; Sherrill, C. D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6106–6114.

(71) Hohenstein, E. G.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 878–886.

(72) Peverati, R.; Baldridge, K. *J. Chem. Theory Comput.* **2008**, *4*, 2030–2048.

CT900551Z

JCTC Journal of Chemical Theory and Computation

Metal−Metal Quintuple and Sextuple Bonding in Bent Dimetallocenes of the Third Row Transition Metals

Bing Xu,[†,‡] Qian-Shu Li,*[,‡,§] Yaoming Xie,[||] R. Bruce King,*[,§,||] and Henry F. Schaefer III[||]

Beijing University of Posts and Telecommunication, Beijing 100876, China, Institute of Chemical Physics, Beijing Institute of Technology, Beijing 100081, China, Center for Computational Quantum Chemistry, South China Normal University, Guangzhou, 510631 China, and Department of Chemistry and Center for Computational Chemistry, University of Georgia, Athens, Georgia 30602

Received October 23, 2009

Abstract: Theoretical studies on the dimetallocenes Cp_2M_2 (M = Os, Re, W, Ta) predict bent structures with short metal−metal distances suggesting high-order metal−metal multiple bonds. Analysis of the frontier bonding molecular orbitals indicates a formal Os−Os quintuple bond ($\sigma + 2\pi + 2\delta$) in singlet Cp_2Os_2 and a formal Re−Re sextuple bond ($2\sigma + 2\pi + 2\delta$) in singlet Cp_2Re_2, thereby giving the metals in both molecules the favored 18-electron metal configurations. Predicted low-energy triplet structures for Cp_2M_2 (M = Os, Re) have formal quintuple bonds but with only two δ one-electron "half" bonds (M = Os) or a single δ two-electron bond (M = Re) and a second σ component derived from overlap of the d(z^2) orbitals. A quintuple bond similar to that found in triplet Cp_2Re_2 is found in singlet Cp_2W_2, giving both tungsten atoms a 16-electron configuration. The formal Ta−Ta quadruple bond in the lowest energy singlet Cp_2Ta_2 structure is different from that in the original $Re_2Cl_8^{2-}$ in that it is a $2\sigma + 2\pi$ bond with no δ components but only σ and π components.

1. Introduction

The chemistry of metal−metal multiple bonding[1,2] dates back to the pioneering work of Cotton and Harris[3] in 1965 on the rhenium−rhenium quadruple bond in the binuclear metal halide complex $Re_2Cl_8^{2-}$. This was not only the first example of a metal−metal quadruple bond but also the first example of a quadruple bond of any type. The highest known formal metal−metal bond order in a stable molecule then remained four for 40 years until the 2005 discovery by Power et al.[4] of a binuclear chromium(I) aryl of the type RCrCrR, with an extremely short metal−metal distance, suggesting a formal quintuple bond. This seminal discovery stimulated numerous theoretical studies on high order metal−metal bonds.[5−11] In addition, various research groups reported further experimental work on low oxidation state transition metal aryls of the type RMMR[12,13] as well as chromium(I) amidinate,[14,15] 2-aminopyridine,[16] and diazadiene[17] complexes, apparently containing formal quintuple bonds.

Another key development in organometallic chemistry in recent years has been the synthesis of formal Zn(I) derivatives with direct Zn−Zn bonds.[18,19] Such compounds include dizincocene, CpZn−ZnCp ($Cp = \eta^5$-C_5H_5), in which two CpZn units are linked by a direct zinc−zinc single bond of length 2.305 Å to give both zinc atoms the favored 18-electron configuration.[18] Simple electron counting guided by the 18-electron rule suggests that analogous dimetallocenes of earlier transition metals could provide interesting new examples of metal−metal multiple bonding. The stability of $Re_2Cl_8^{2-}$ with a formal rhenium−rhenium quadruple bond[3] suggests that the best candidates for stable dimetallocenes with interesting metal−metal mul-

* Corresponding author e-mail: rbking@chem.uga.edu (R.B.K.) and qsli@scnu.edu.cn (Q.-S.L.).

[†] Beijing University of Posts and Telecommunication.

[‡] Beijing Institute of Technology.

[§] South China Normal University.

[||] University of Georgia.

10.1021/ct900564p
Published on Web 02/17/2010

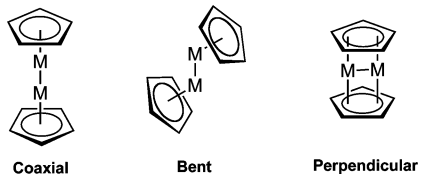


Figure 1. Three types of dimetallocene structures.

tiple bonds might contain the heaviest transition metals, particularly those of the third row.

These considerations led us to investigate dimetallocenes Cp_2M_2 of the early transition metals of the third row (M = Os, Re, W, Ta) as candidates for new types of molecules with high order metal−metal bonds. In this connection, all of these dimetallocenes were found by density functional theory (DFT) to exhibit bent structures in contrast to the coaxial structure found[18] for Cp_2Zn_2 and the perpendicular structures for late transition metal metallocenes predicted by theory in 2005 (Figure 1).[20]

The theoretical studies of the dimetallocenes Cp_2M_2 (M = Os, Re, W, Ta) discussed in this paper consist of two phases: (1) optimizations of structures with singlet, triplet, and quintet spin states using density functional methods and (2) elucidation of the formal metal−metal bond orders in the lowest energy optimized structures by analysis of the highest occupied molecular orbitals (MOs). The MO studies provide more direct evidence for the high order metal−metal multiple bonds, already suggested by the relatively short metal−metal distances. Thus, evidence is provided for the formal osmium−osmium quintuple bond in Cp_2Os_2 required to give both osmium atoms the favored 18-electron rare gas configuration.

2. Theoretical Methods

Electron correlation effects were considered by employing density functional theory (DFT), which has evolved as a practical and effective computational tool, especially for organometallic compounds.[21−35] In this research, two DFT methods, BP86 and MPW1PW91, were used. The BP86 method is a pure DFT method that combines Becke's 1988 exchange functional with Perdew's 1986 correlation functional.[36,37] The MPW1PW91 method[38] is a so-called second generation[39] functional, which combines the modified Perdew−Wang exchange functional with Perdew−Wang's 1991 correlation functional.[40] The MPW1PW91 method has been found to be typically more suitable for geometry optimization of the second and third row transition metal systems,[41,42] while the BP86 method usually provides better vibrational frequencies with DZP basis sets.

For the third row transition metals, the large numbers of electrons may increase exponentially the computational efforts. In order to reduce the cost, effective core potential (ECP) relativistic basis sets are employed. The SDD (Stuttgart−Dresden ECP plus DZ)[43] ECP basis set was used for the Os, Re, W, and Ta atoms. For the C atom, the double-ζ plus polarization (DZP) basis set was used. The latter are the Huzinaga−Dunning contracted double-ζ sets[44,45] plus a set of spherical harmonic d polarization functions with an orbital exponent $\alpha_d(C) = 0.75$, designated as (9s5p1d/4s2p1d). For H, a set of p polarization functions, $\alpha_p(H) = 0.75$, was added to the Huzinaga−Dunning DZ set.

The geometries of all structures were fully optimized using the two selected DFT methods with the SDD ECP basis set. The vibrational frequencies were determined by evaluating analytically the second derivatives of the energy with respect to the nuclear coordinates at the same theoretical levels. The corresponding infrared intensities were also evaluated analytically.

All of the computations were carried out with the Gaussian 03 program.[46] The fine (75, 302) grid is the default for evaluating integrals numerically, and the tight (10^{-8} hartree) designation is the default for the energy convergence. The finer grid (120, 974) was used for more carefully characterizing small imaginary vibrational frequencies.

3. Results

3.1. Cp_2M_2 (M = Os, Re, W, Ta) Structures. *3.1.1. Cp_2Os_2.* Initially, we investigated the linear D_{5h} and D_{5d} structures of Cp_2Os_2, which are similar to the experimental Cp_2Zn_2 structure.[18] However, these structures were found to have four imaginary vibrational frequencies at 160*i*, 160*i*, 56*i*, and 56*i* cm^{-1} (for D_{5h}) or 160*i*, 160*i*, 56*i*, and 56*i* cm^{-1} (for D_{5d}). Following the corresponding normal modes leads to the bent C_{2h} structure Os−S, which lies below the two linear structures by ∼49 kcal/mol.

The bent C_{2h} singlet structure Os−S (Figure 2 and Table 1) is the global minimum. The Os−Os distance in Os−S is predicted to be 2.333 Å (MPW1PW91) or 2.276 Å (BP86). This is short enough to correspond to the formal quintuple bond required to give both osmium atoms the favored 18-electron configuration. Furthermore, the Os−Os distance in Os−S is ∼0.5 Å shorter than the experimental Os−Os single bond distance of 2.767 Å in $(\eta^5\text{-}Me_5C_5)_2Os_2(CO)_2(\mu\text{-}CO)_2$ determined by X-ray crystallography.[47]

The triplet structure Os−T for Cp_2Os_2 (Figure 2) is predicted to lie only 0.1 kcal/mol (MPW1PW91) lower or 3.6 kcal/mol (BP86) higher in energy than the singlet structure Os−S (Table 1). A small imaginary vibrational frequency at 67*i* cm^{-1} is predicted by the MPW1PW91 method, while all real vibrational frequencies are predicted by the BP86 method. The 67*i* cm^{-1} imaginary vibrational frequency cannot be removed by using a finer integration grid (120, 974). Following the corresponding normal mode of the imaginary vibrational frequency leads to a C_1 structure, which only slightly deviates from the C_2 structure Os−T. The Os−Os bond length in Os−T is predicted to be 2.277 Å (MPW1PW91) or 2.255 Å (BP86).

The quintet structure Os−Q for Cp_2Os_2 (C_2) is predicted to have all real vibrational frequencies by both DFT methods and lies 6.7 kcal/mol (MPW1PW91) or 17.0 kcal/mol (BP86) higher in energy than the singlet structure Os−S (Table 1). The Os−Os bond length in Os−Q is predicted to be essentially the same as Os−T, namely, 2.281 Å (MPW1PW91) or 2.290 Å (BP86), respectively. Note that all three Os−Os distances are essentially the same for the three Cp_2Os_2 structures Os−S, Os−T, and Os−Q regardless of the spin state.

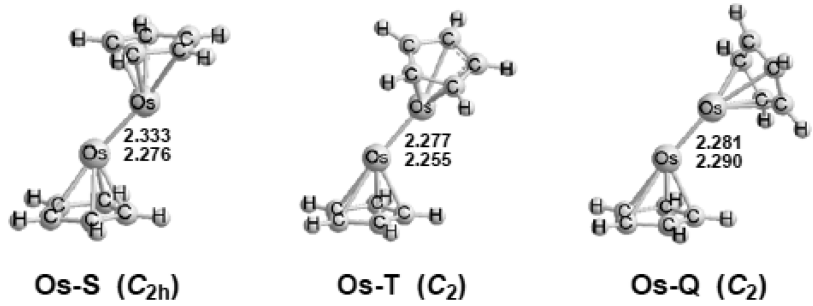


Figure 2. Optimized structures for Cp_2Os_2. Distances are reported in Å. The upper distances were predicted by the MPW1PW91 method and the lower distances by the BP86 method (same for subsequent figures).

Table 1. Total Energies (E, in Hartree), Relative Energies (ΔE, in kcal/mol), Numbers of Imaginary Vibrational Frequencies (Nimag), Os−Os Bond Distances (Å), and Spin Expectation Values $\langle S^2 \rangle$ for the Optimized Cp_2Os_2 Structures

		Os−S (C_{2h})	Os−T (C_2)	Os−Q (C_2)
	state	^1A_g	^3B	^5A
MPW1PW91	E	−568.42221	−568.42243	−568.41160
	ΔE	0.0	−0.1	6.7
	Nimag	0	1(67i)	0
	Os−Os	2.333	2.277	2.281
	$\langle S^2 \rangle$	0	2.02	6.03
BP86	E	−568.71958	−568.71385	−568.69253
	ΔE	0.0	3.6	17.0
	Nimag	0	0	0
	Os−Os	2.276	2.255	2.290
	$\langle S^2 \rangle$	0	2.01	6.02

3.1.2. Cp_2Re_2. The global minimum of Cp_2Re_2 is predicted to be a C_2 triplet *trans* structure Re−T_1 (Figure 3 and Table 2) with all real vibrational frequencies by the MPW1PW91 method but with a very small imaginary vibrational frequency at 18i cm^{-1} by the BP86 method. This imaginary vibrational frequency is removed by using a finer integration grid (120, 974), indicating that it is caused by numerical integration error. The Re−Re distance in Re−T_1 is predicted to be very short at 2.323 Å (MPW1PW91) or 2.272 Å (BP86), suggesting a metal−metal bond of high multiplicity. In this connection, the formal Re≡Re triple bond in (η^5-Me_5C_5)$_2$Re$_2$(μ-CO)$_3$ is found experimentally by X-ray diffraction[48] to be 2.411 Å, suggesting on the basis of bond length a formal bond order appreciably greater than three for the appreciably shorter Re−Re bond in Re−T_1. A related *cis* C_2 triplet Cp_2Re_2 structure Re−T_2 (Figure 3 and Table 2) is also predicted to be a genuine minimum, lying only 0.9 kcal/mol (MPW1PW91) or 2.3 kcal/mol (BP86) above the global minimum Re−T_1.

A singlet Cp_2Re_2 structure Re−S (Figure 3 and Table 2) is predicted at 8.0 kcal/mol (MPW1PW91) or 0.6 kcal/mol (BP86) in energy above Re−T_1 with all real vibrational frequencies. The Re−Re distance in Re−S is predicted to be even 0.08 Å shorter than the already very short Re−Re distance in Re−T_1. This is consistent with the requirement in Cp_2Re_2 of a formal quintuple bond to give the rhenium atoms in Cp_2Re_2 the 17-electron rhenium configurations in the binuclear triplet Re−T_1 but a formal sextuple bond for the favored 18-electron rhenium configurations in the singlet Re−S.

The C_{2h} symmetry quintet Cp_2Re_2 structure Re−Q_3 (Figure 3 and Table 2) is predicted to lie 2.3 kcal/mol (MPW1PW91) or 8.8 kcal/mol (BP86) higher in energy than the global minimum Re−T_1. The MPW1PW91 method predicts all real vibrational frequencies for Re−Q_3(C_{2h}), while a substantial imaginary vibrational frequency at 142i cm^{-1} is predicted by the BP86 method. Following the corresponding normal mode of this imaginary vibrational frequency leads to structure Re−Q_1 (C_i, Figure 3), which is 1.2 kcal/mol (BP86) lower in energy than Re−Q_3 (Table 2). A *cis* Cp_2Re_2 quintet C_2 structure Re−Q_2 (Figure 3 and Table 2) is predicted to lie 0.9 kcal/mol (BP86) higher in energy than Re−Q_1 (C_i). The Re−Re bond lengths in Re−Q_1 and Re−Q_2 are almost the same at 2.305 Å (BP86), whereas a slightly longer Re−Re bond at 2.339 Å (BP86) is predicted for structure Re−Q_3. Since unstable electronic states could be the computational results for the bimetallic systems and Re−Q_1 and Re−Q_2 are two prime candidates for this problem, we checked the stability for these two structures. However, our results confirm that they are both stable.

3.1.3. Cp_2W_2. The C_2 singlet Cp_2W_2 structure W−S (Figure 4 and Table 3) is predicted to be the global minimum of Cp_2W_2 by the BP86 method. The triplet and quintet structures of Cp_2W_2 are predicted by BP86 to lie 9.0 and 10.3 kcal/mol, respectively, higher in energy than W−S (Table 3). However, the C_2 quintet structure W−Q is predicted to be the global minimum for Cp_2W_2 by MPW1PW91 at just 1.4 kcal/mol lower in energy than W−S. All four Cp_2W_2 structures are found to have all real vibrational frequencies by BP86. Whereas, the Cp_2W_2 structures W−S and W−T_1 (Figure 4, Table 3) are predicted to have small imaginary vibrational frequencies at 22i cm^{-1} (W−S) or 32i cm^{-1} (W−T_1) with the MPW1PW91 method (Table 3). These small imaginary vibrational frequencies are not removed by the finer integration grid of (120, 974). Following the corresponding normal modes (Cp ring rotations) of these small imaginary vibrational frequencies led to C_1 structures with essentially unchanged W−W distances and energies lower by only ∼1 kcal/mol. The triplet structures W−T_1 and W−T_2 (Figure 4, Table 3) exhibit significant spin contamination by the MPW1PW91 method, namely, $\langle S^2 \rangle = 2.33$ and 2.23 versus the ideal $S(S + 1) = 2$, however, the BP86 method predicts less spin contamination ($\langle S^2 \rangle = 2.07$ and 2.09) for these two triplet structures. The W−W distances of Cp_2W_2 structures are predicted to fall in the range of 2.291 Å to 2.414 Å, consistent with higher order W−W

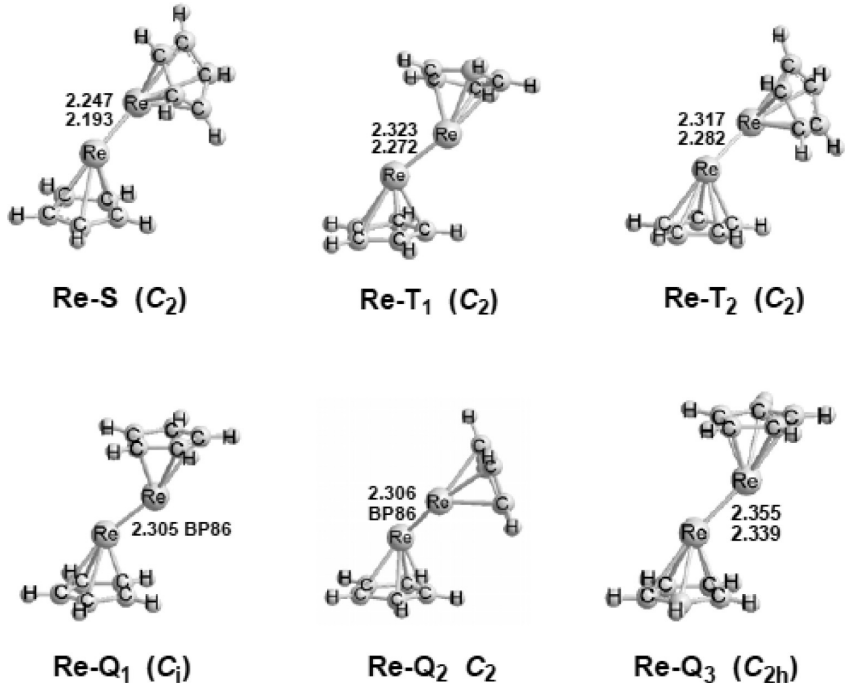


Figure 3. Optimized structures of Cp_2Re_2.

Table 2. Total Energies (E, in Hartree), Relative Energies (ΔE, in kcal/mol), Numbers of Imaginary Vibrational Frequencies (Nimag), Re−Re Bond Distances (Å), and Spin Expectation Values $\langle S^2 \rangle$ for the Optimized Cp_2Re_2 Structures

		Re−S	Re−T$_1$	Re−T$_2$	Re−Q$_1$	Re−Q$_2$	Re−Q$_3$
		(C_2)	(C_2)	(C_2)	(C_i)	(C_2)	(C_{2h})
	state	^{1}A	^{3}B	^{3}B	^{5}A$_g$	^{5}A	^{5}B$_g$
MPW1PW91	E	−543.56593	−543.57860	−543.57723			−543.57497
	ΔE	8.0	0.0	0.9			2.3
	Nimag	0	0	0	same as Re−Q$_3$ (C_{2h})		0
	Re−Re	2.247	2.323	2.317			2.355
	$\langle S^2 \rangle$	0	2.22	2.03			6.08
BP86	E	−543.85837	−543.85931	−543.84719	−543.84579	−543.84579	−543.84533
	ΔE	0.6	0.0	2.3	7.6	8.5	8.8
	Nimag	0	1(18i)	0	0	0	1(142i)
	Re−Re	2.193	2.272	2.282	2.305	2.306	2.339
	$\langle S^2 \rangle$	0	2.04	2.01	6.03	6.08	6.03

multiple bonds. For comparison, the W≡W triple bond distance found by X-ray crystallography[49] in $Cp_2W_2(CO)_2(\mu$-$Ph_2PCH_2PPh_2)$ is 2.514 Å. This suggests formal bond orders greater than three for the Cp_2W_2 structures.

3.1.4. Cp_2Ta_2. The global minimum of Cp_2Ta_2 is predicted to be a C_2 triplet structure Ta−T, having all real vibrational frequencies (Figure 5, Table 4) and a Ta−Ta bond distance of 2.467 Å (MPW1PW91) or 2.480 Å (BP86). In addition, *trans* and *cis* singlet Cp_2Ta_2 structures Ta−S$_1$ (C_{2h}) and Ta−S$_2$ (C_2) are predicted to lie above Ta−T by ∼5.7 kcal/mol or ∼11.2 kcal/mol by MPW1PW91 but ∼0.7 kcal/mol lower or ∼4.9 kcal/mol higher in energy by BP86 (Figure 5, Table 4). These two singlet structures are both predicted to be genuine minima, having all real vibrational frequencies. The short Ta−Ta bond lengths at 2.369 Å to 2.417 Å in the singlet Cp_2Ta_2 structures are again consistent with higher order multiple bonds.

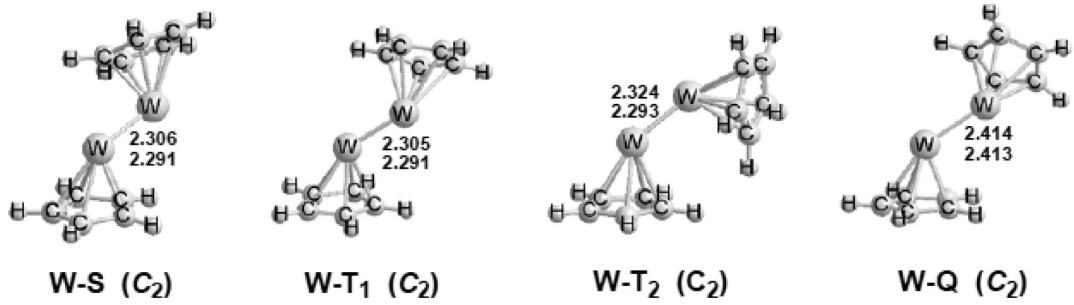


Figure 4. Optimized structures of Cp_2W_2.

Table 3. Total Energies (*E*, in Hartree), Relative Energies (Δ*E*, in kcal/mol), Numbers of Imaginary Vibrational Frequencies (Nimag), W−W Bond Distances (Å), and Spin Expectation Values $\langle S^2 \rangle$ for the Optimized Cp_2W_2 Structures

		W−S (C_2)	W−T_1 (C_2)	W−T_2 (C_2)	W−Q (C_2)
	state	^{1}A	^{3}B	^{3}B	^{5}A
MPW1PW91	*E*	−521.13768	−521.12591	−521.12455	−521.13991
	Δ*E*	0.0	7.4	8.2	−1.4
	Nimag	1(22*i*)	1(32*i*)	0	0
	W−W	2.306	2.305	2.324	2.414
	$\langle S^2 \rangle$	0	2.33	2.23	6.05
BP86	*E*	−521.40576	−521.39143	−521.38738	−521.38932
	Δ*E*	0.0	9.0	11.5	10.3
	Nimag	0	0	0	0
	W−W	2.291	2.291	2.293	2.413
	$\langle S^2 \rangle$	0	2.07	2.09	6.02

The *trans* and *cis* quintet structures Ta−Q_1 (C_{2h}) and Ta−Q_2 (C_1), respectively, are predicted to have all real vibrational frequencies by BP86 (Figure 5, Table 4), while a small imaginary vibrational frequency at 29*i* cm^{-1} for Ta−Q_1 is predicted by MPW1PW91. Following the corresponding normal mode, which corresponds to Cp ring rotation, leads to a C_s structure with an energy lower by only ∼0.2 kcal/mol. The two quintet Cp_2Ta_2 structures are predicted to lie in energy above Ta−T by ∼6.0 kcal/mol (MPW1PW91) or ∼10.5 kcal/mol (BP86). The Ta−Ta distances in Ta−Q_1 and Ta−Q_2 are predicted to fall in the range of 2.600 Å to 2.664 Å, which is ∼0.15 to ∼0.25 Å longer than those for the singlet and triplet Cp_2Ta_2 structures.

3.2. Cp_2M_2 (M = Os, Re, W, Ta) Molecular Orbitals. The very short M−M bond distances in the Cp_2M_2 derivatives (M = Os, Re, W, Ta) suggest metal−metal multiple bonds of relatively high bond orders for practically all of the predicted structures. For singlet Cp_2Os_2 and singlet Cp_2Re_2, these metal−metal distances are even short enough to suggest the formal quintuple and sextuple bonds, respectively, required to give both metal atoms the favored 18-electron configurations. In order to obtain further evidence beyond simply very short M−M distances for these very interesting metal−metal multiple bonds in the Cp_2M_2 derivatives, their frontier molecular orbitals (MOs) were investigated. The highest occupied molecular orbitals (HOMO) as well as the (typically) five or six orbitals below the HOMO were found to be localized mainly on the metal−metal bonds with relatively little on the Cp rings. These frontier MOs were used to characterize the metal−metal bonding in the dimetallocenes Cp_2M_2 (M = Os, Re, W, Ta). In this connection, the formal bond order of the metal−metal bond (FBO_{M-M}) can be defined by the equation $FBO_{M-M} = {}^1/_2(n_B - n_A)$ where n_B and n_A are the numbers of electrons in the bonding and antibonding orbitals, respectively. The analyses for the all of the optimized Cp_2M_2 structures discussed in this paper are summarized in Table 5, along with the M−M bond distances determined by the BP86 method, where data for all of the structures are available. The MOs for the lowest energy Cp_2M_2 structures exhibiting metal−metal quintuple

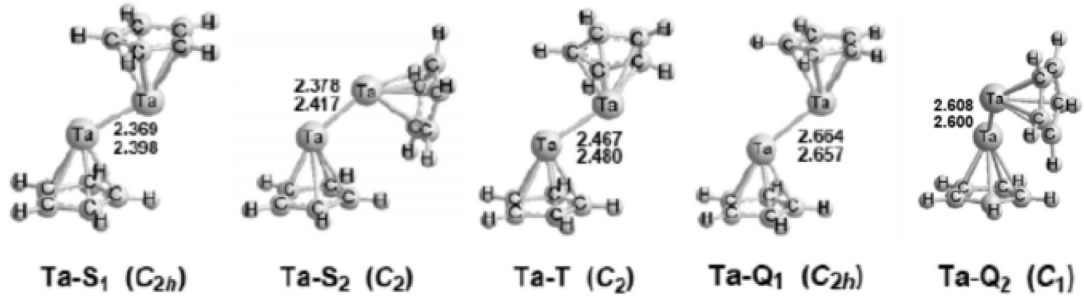


Figure 5. Optimized structures of Cp_2Ta_2.

Table 4. Total Energies (*E*, in Hartree), Relative Energies (Δ*E*, in kcal/mol), Numbers of Imaginary Vibrational Frequencies (Nimag), Ta−Ta Bond Distances (Å), and Spin Expectation Values $\langle S^2 \rangle$ for the Optimized Cp_2Ta_2 Structures

		Ta−S_1 (C_{2h})	Ta−S_2 (C_2)	Ta−T (C_2)	Ta−Q_1 (C_{2h})	Ta−Q_2 (C_1)
	state	^1A_g	^{1}A	^{3}B	^5A_g	^{5}A
MPW1PW91	*E*	−500.97396	−500.96438	−500.98219	−500.97303	−500.97233
	Δ*E*	5.7	11.2	0.0	5.7	6.2
	Nimag	0	0	0	1(29*i*)	0
	Ta−Ta	2.369	2.378	2.467	2.664	2.608
	$\langle S^2 \rangle$	0	0	2.01	6.02	6.08
BP86	*E*	−501.21582	−501.20683	−501.21469	−501.19919	−501.19673
	Δ*E*	−0.7	4.9	0.0	9.7	11.3
	Nimag	0	0	0	0	0
	Ta−Ta	2.398	2.417	2.480	2.657	2.600
	$\langle S^2 \rangle$	0	0	2.01	6.01	6.02

Table 5. Formal Metal−Metal Bond Order in the Dimetallocene Structures Cp_2M_2 (M = Os, Re, W, Ta) as Determined by an Analysis of the Bonding Molecular Orbitals

structure	bonding electrons	antibonding electrons	M−M bond order[a] (FBO_{M-M})	M−M distance, Å (BP86)
		Cp_2Os_2		
Os−S	12	2	5	2.276
Os−T	12	2	$4^2/_2$	2.255
Os−Q	12	2	5	2.290
		Cp_2Re_2		
Re-S	12	0	6	2.193
$Re-T_1$	11	1	5	2.272
$Re-T_2$	11	1	5	2.282
$Re-Q_1$	10	2	4	2.305
$Re-Q_2$	10	2	4	2.306
$Re-Q_3$	10	2	4	2.339
		Cp_2W_2		
W−S	10	0	5	2.291
$W-T_1$	10	0	$4^2/_2$	2.291
$W-T_2$	10	0	$4^2/_2$	2.293
W−Q	9	1	4	2.413
		Cp_2Ta_2		
$Ta-S_1$	8	0	4	2.398
$Ta-S_2$	8	0	4	2.417
Ta−T	7	1	3	2.480
$Ta-Q_1$	7	1	3	2.657
$Ta-Q_2$	7	1	3	2.600

[a] The notation "$^2/_2$" refers to a pair of essentially degenerate one-electron "half bonds."

and sextuple bonding are discussed in detail below. A complete set of the figures of the relevant frontier orbitals of all of the optimized Cp_2M_2 structures is given in the Supporting Information.

The M−M bond distances listed in Table 5 show a rough correlation with the formal metal−metal bond order FBO_{M-M} as determined by the numbers of electrons in M−M bonding and M−M antibonding orbitals. Thus, the unique structure with a formal sextuple bond, namely, the singlet Cp_2Re_2 structure Re−S, has a Re−Re bond distance of 2.193 Å, which is shorter than any of the other Cp_2M_2 derivatives by 0.06 Å or more. The Cp_2M_2 derivatives with formal M−M quintuple bonds are predicted to have metal−metal distances in the rather narrow range of 2.25 to 2.29 Å. Comparison of these M−M distances with the 2.222 Å length of the Re−Re quadruple bond in $Re_2Cl_8^{2-}$ found experimentally[50] suggests that Cp ligands lead to longer M−M bonds of a given multiplicity than chloride ligands. The formal M−M quadruple bonds in the Cp_2M_2 derivatives are predicted to fall in the range 2.30 to 2.41 Å and thus are significantly longer than the formal Re−Re quadruple bond in $Re_2Cl_8^{2-}$. The Ta≡Ta triple bonds in the Cp_2Ta_2 derivatives Ta−T, $Ta-Q_1$, and $Ta-Q_2$ are still longer at 2.48 to 2.66 Å.

The method of determining the formal bond orders is illustrated for Cp_2Os_2 by the singlet structure Os−S (Figures 2 and 6). In this case, the seven highest energy occupied MOs (HOMO down to HOMO−6) have their electron densities concentrated on the metals rather than the rings and thus may be assumed to be responsible for the metal−metal bonding. The lower occupied MOs have most of their electron density on the rings (HOMO−7 to HOMO−10 in Figure 6 and further down) and thus may be related to metal−ring bonding. Among the seven highest occupied MOs, one (HOMO) is clearly antibonding and the other six (HOMO−1 down to HOMO−6) are clearly bonding since

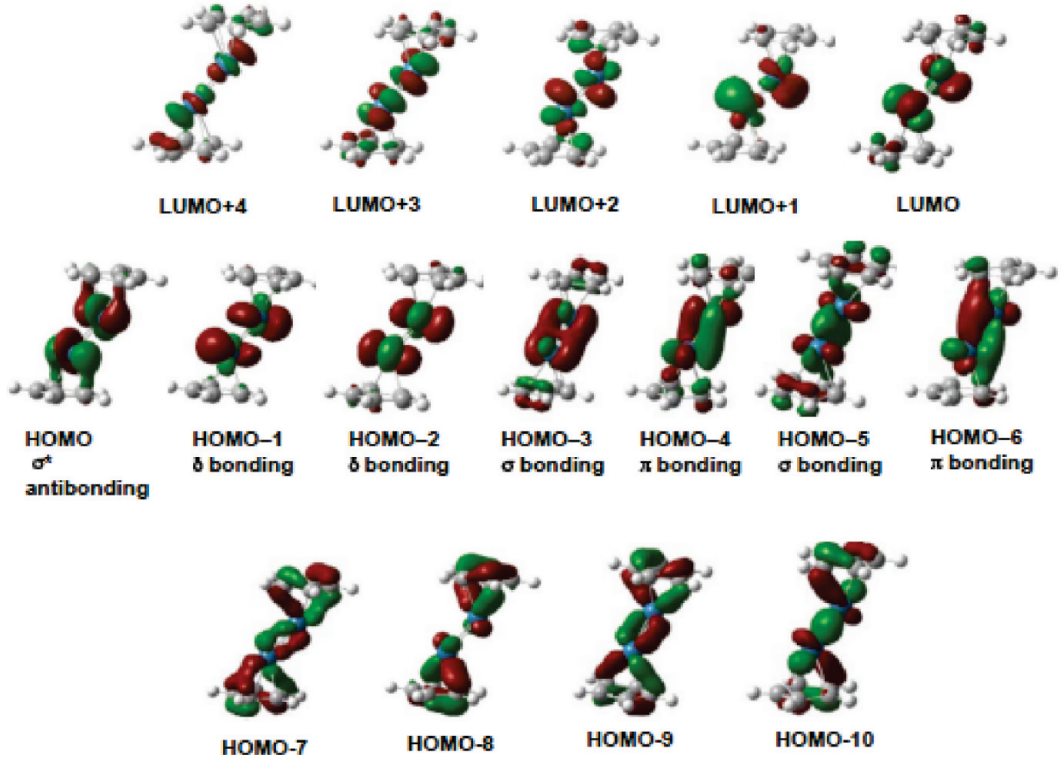


Figure 6. The frontier MOs for the singlet structure Os−S of Cp_2Os_2. (a) Top: The unoccupied LUMOs up to LUMO+4. (b) Middle: The occupied HOMOs down to HOMO−6 relating to the osmium−osmium bonding. (c) Bottom: The next lower occupied MOs (HOMO−7 down to HOMO−10), relating largely to the metal−ring bonding.

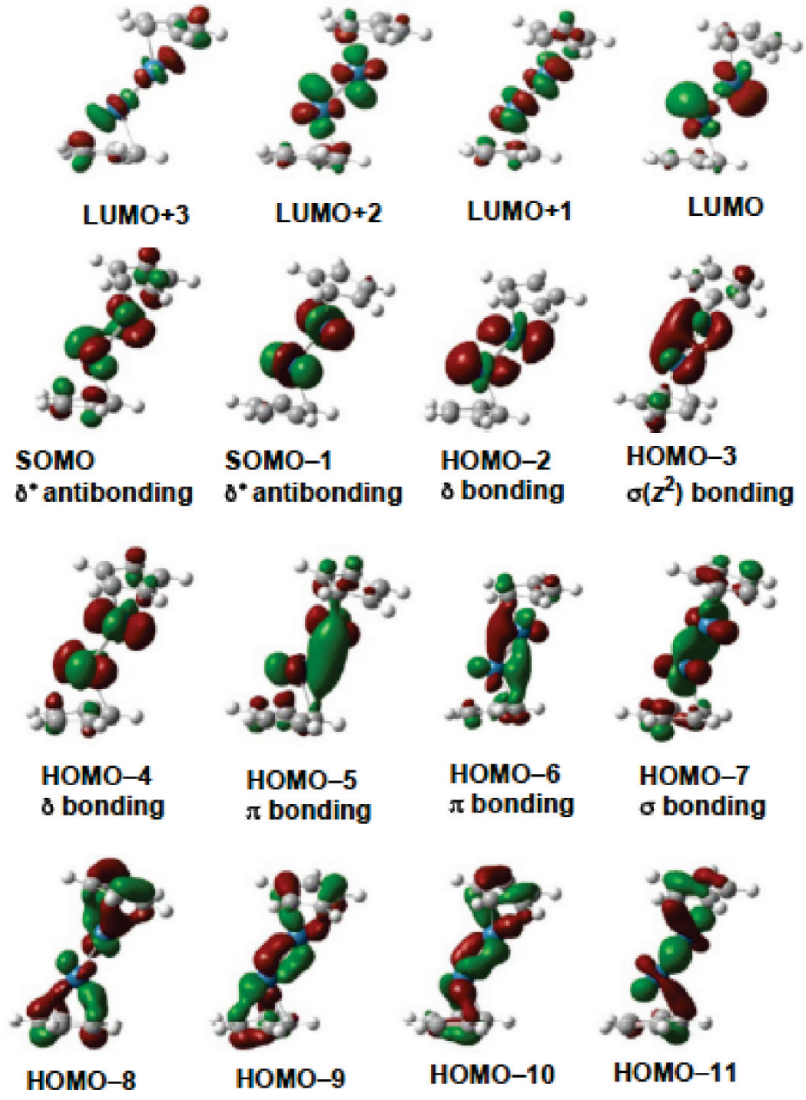


Figure 7. The frontier MOs for the triplet structure Os−T of Cp_2Os_2. (a) LUMOs up to LUMO+3 are unoccupied orbitals. (b) SOMO and SOMO−1 are osmium−osmium antibonding orbitals. (c) HOMO−2 to HOMO−7 are osmium−osmium bonding orbitals. (d) Bottom: The next lower occupied MOs (HOMO−8 down to HOMO−11) relate largely to the metal−ring bonding.

there is no node between the metal atoms. This pattern of the occupied MOs (six bonding orbitals and one antibonding orbital) leads to a formal bond order of $^1/_2(12 - 2) = 5$, i.e., the formal quintuple bond needed to give both osmium atoms the favored 18-electron configuration in Cp_2Os_2.

The bent nature of the diosmocene structure Os−S (Figure 2) makes less clear the nature of the components of the osmium−osmium quintuple bond. However, the six bonding orbitals HOMO−1 to HOMO−6 (Figure 6) appear to correspond to two δ bonds, a $\sigma(z^2)$ bond, a π bond, a σ(s) bond, and another π bond, respectively. The filled antibonding orbital (HOMO) appears to be a $\sigma^*(z^2)$ orbital, thereby canceling out the $\sigma(z^2)$ bonding component and leaving the five σ(s) + 2π + 2δ bonding components for the quintuple bond, similar to the quintuple bond in the binuclear Cr(I) aryl derivative RCrCrR of Power et al.[4] Also, the shapes of the MOs for Cp_2Os_2 (Figure 6) indicate the expected weaker overlap in the two δ components HOMO−1 and HOMO−2 relative to the σ and π components.

A low energy triplet structure Os−T (Figure 2) is also found for Cp_2Os_2. The frontier molecular orbitals for Os−T are shown in Figure 7. The two MOs below the LUMO, namely, SOMO and SOMO−1, contain only a single electron corresponding to the triplet spin multiplicity of Os−T. Both of these orbitals correspond to δ^* antibonding orbitals. The next six orbitals below SOMO−1, namely, HOMO−2 through HOMO−7, each contain electron pairs and correspond to a δ bonding orbital, a $\sigma(z^2)$ bonding orbital, another δ bonding orbital, two π bonding orbitals, and a σ bonding orbital. Thus, the eight orbitals from SOMO down to HOMO−7 correspond to an Os−Os quintuple bond consisting of $2\sigma + 2\pi$ full bonds and two δ half bonds, i.e., a bond of order $4^2/_2$. This is analogous to the well-known bond of order $1^2/_2$ in normal (triplet) dioxygen except that in dioxygen the half bonds are π bonds rather than the δ bonds in structure Os−T of Cp_2Os_2. Thus, both singlet Cp_2Os_2 (Os−S) and triplet Cp_2Os_2 (Os−T) have metal−metal bonds of order 5 to give the osmium atoms the favored 18-electron configurations. In the singlet Os−S, the Os−Os quintuple bond is of the type $\sigma + 2\pi + 2\delta$ with one σ bond and two full two-electron δ bonds. However, in triplet Os−T, the Os−Os quintuple bond is of the type $2\sigma + 2\pi + {}^2/_2\delta$

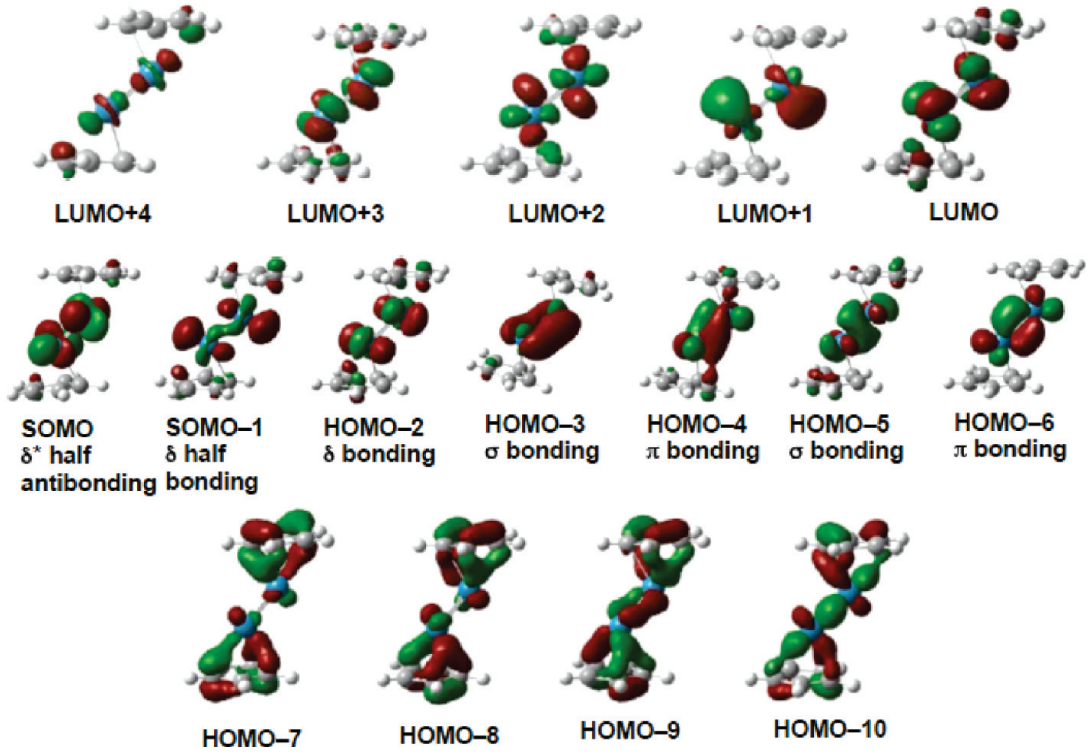


Figure 8. The frontier MOs for structure Re−T_1 of Cp_2Re_2. (a) Top: The unoccupied LUMO up to LUMO+4. (b) Middle: The occupied orbitals SOMO, SOMO−1, and HOMO−2 down to HOMO−6, relating to the rhenium−rhenium bonding. (c) Bottom: The next lower occupied MOs (HOMO−7 down to HOMO−10), relating largely to the metal−ring bonding.

with two δ single electron "half" bonds and a second σ bond based on overlap of the d(z^2) orbitals. The predicted Os−Os distances for these two types of quintuple bonds are within 0.02 Å of each other, namely, 2.276 Å for the singlet Os−S and 2.255 Å for the triplet Os−T.

Dirhenocene, Cp_2Re_2, has two electrons less than disomocene Cp_2Os_2. In the singlet structure Re−S (Figure 3) for dirhenocene, the σ* antibonding HOMO of Cp_2Os_2 (Figure 6) is empty because of these "missing" two electrons. Therefore, the Re−Re bond in Re−S has six components, namely, two π components and four σ and δ components. Because of the bending in Re−S, the σ and δ components are not readily distinguishable. These six components of the Re−Re bond in Re−S imply that the sextuple bond required to give both rhenium atoms the favored 18-electron configuration. Thus, six of the nine orbitals in the sp^3d^5 manifolds of each rhenium atom are allocated to the Re−Re sextuple bond leaving three orbitals on each metal atom for the $\sigma + 2\pi$ components of the metal−ring bonds. A sextuple bond has been postulated for the bare dimers M_2 (M = Cr, Mo, W) of the group 6 metals.[51] Note that the d^6 formal Re(I) in Cp_2Re_2 is isoelectronic with the d^6 formal M(0) in the group 6 metal dimers (considering the Cp ring as Cp^- with the favorable 6 π electrons). The formal sextuple bond in the Cp_2Re_2 structure Re−S is consistent with its predicted 2.22 ± 0.03 Å distance being ~0.08 Å shorter than the formal Os−Os quintuple bond distance of 2.30 ± 0.03 Å in the singlet Cp_2Os_2 structure Os−S.

The lowest energy Cp_2Re_2 structure by either method is not the singlet Re−S but the triplet Re−T_1, which lies marginally lower in energy (Figure 3 and Table 2). The frontier MOs of Re−T_1 are shown in Figure 8. The Re−T_1 SOMOs, like those in the triplet Cp_2Os_2 structure Os−T discussed above, contain only a single electron, yielding triplet spin multiplicity. One of these half-filled orbitals (SOMO) is a δ* antibonding orbital, whereas the other half-filled orbital (SOMO−1) is a δ bonding orbital. Thus, these two half-filled orbitals in Re−T_1 make no net contribution to the rhenium−rhenium bonding. The five orbitals below the SOMOs, namely HOMO−2 to HOMO−6, inclusive (Figure 8), all contain electron pairs and have most of their electron density on the rhenium atoms rather than the Cp rings. These five orbitals represent the five components of a Re−Re quintuple bond thereby giving both rhenium atoms the 17-electron configurations for a binuclear triplet. However, this Re−Re quintuple bond in Re−T_1 has $\sigma(s) + \sigma(z^2) + 2\pi + \delta$ components rather than the $\sigma(s) + 2\pi + 2\delta$ components of the Os−Os quintuple bond in Os−S (Figure 6). The formal quintuple bond in the Cp_2Re_2 triplet Re−T_1 is consistent with the predicted Re−Re distance of 2.30 ± 0.03 Å, which is essentially identical to the Os−Os quintuple bond distance in Os−S and longer than the Re−Re sextuple bond distance of 2.22 ± 0.03 Å for singlet Re−S. The Re−T_1 orbitals below HOMO−6 (e.g., HOMO−7 through HOMO−10 in Figure 8) have most of their electron density on the Cp rings and thus may be related to the rhenium−ring bonding.

Ditungstenocene, Cp_2W_2, has two electrons less than dirhenocene. The lowest energy Cp_2W_2 structure is the singlet W−S (Figure 4). The frontier MOs of W−S are shown in Figure 9. The five highest lying filled MOs (HOMO down to HOMO−4) have their electron densities concentrated on the metal−metal bond and are all bonding orbitals. This

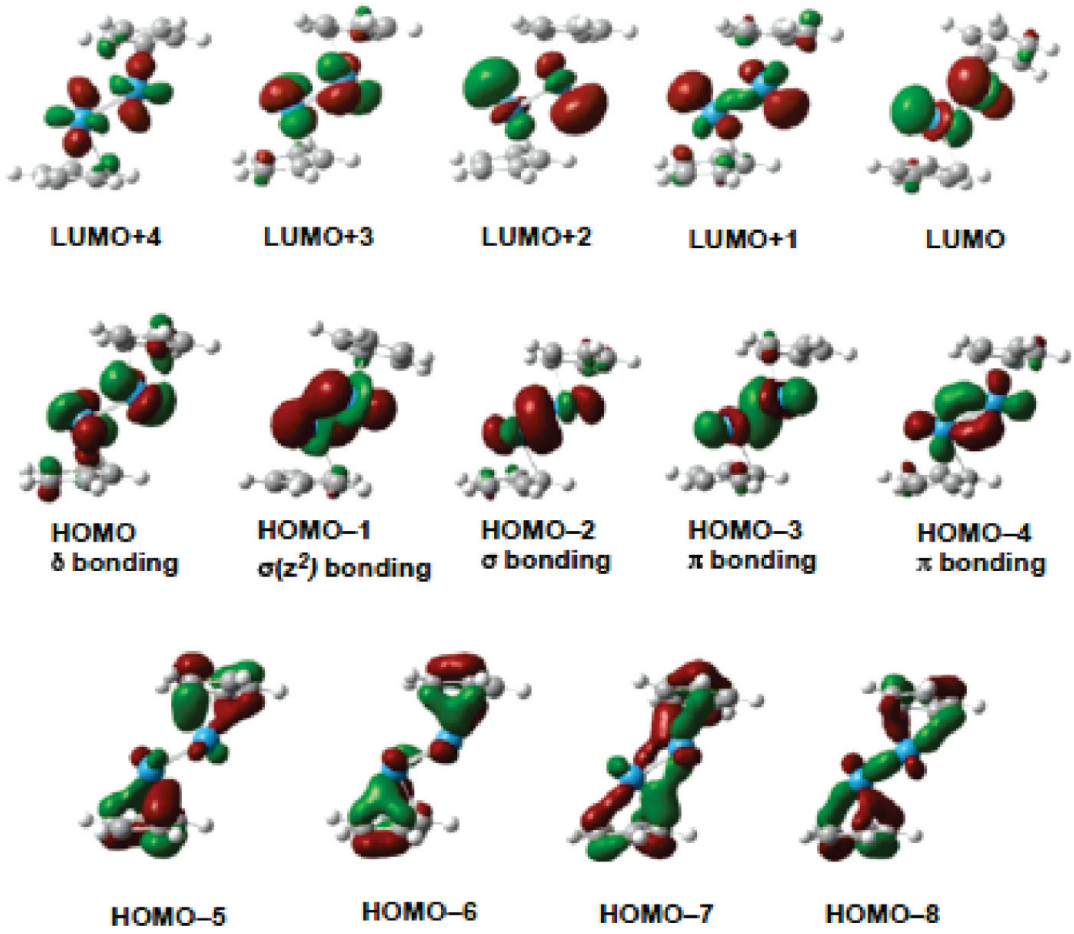


Figure 9. The frontier MOs for structure W−S of Cp_2W_2. (a) Top: The LUMOs up to LUMO+4. (b) Middle: The occupied HOMOs down to HOMO−4, relating to the tungsten−tungsten bonding. (c) Bottom: The next lower occupied MOs (HOMO−5 down to HOMO−8), relating largely to the metal−ring bonding.

indicates a formal quintuple W−W bond in W−S giving both tungsten atoms a 16-electron configuration. This quintuple bond has one δ component (HOMO), two σ components (HOMO−1 and HOMO−2), and two π components (HOMO−3 and HOMO−4) and thus is similar to the Re−Re quintuple bond in the triplet Cp_2Re_2 structure $Re-T_1$. The predicted W−W quintuple bond distance in W−S of 2.30 ± 0.01 Å is essentially identical with the predicted metal−metal quintuple bond distances in the singlet Cp_2Os_2 structure Os−S and the triplet Cp_2Re_2 structure $Re-T_1$.

Ditantalocene, Cp_2Ta_2, has two electrons less than ditungstocene. The frontier MOs of the lowest energy singlet Cp_2Ta_2 structure $Ta-S_1$ are shown in Figure 10. The four highest lying filled MOs (HOMO down to HOMO−4) have their electron densities concentrated on the metal−metal bond and are all bonding orbitals. This indicates a formal quadruple Ta−Ta bond in $Ta-S_1$. This quadruple bond has two σ components (HOMO and HOMO−3) and two π components (HOMO−1 and HOMO−2). Thus, the two electrons lost in going from singlet Cp_2W_2 to singlet Cp_2Ta_2 come from the δ component of the metal−metal multiple bond. The reduction in the formal metal−metal bond order from five in singlet Cp_2W_2 W−S to four in singlet Cp_2Ta_2 $Ta-S_1$ is consistent with a lengthening of the predicted metal−metal distance from 2.30 ± 0.01 Å in W−S to 2.38 ± 0.02 Å in $Ta-S_1$.

4. Discussion

Analysis of the metal−metal multiple bonding in the dimetallocenes uses the 18-electron rule and variations thereof to determine the formal metal−metal multiple bond orders. Thus, the singlet structures of Cp_2Os_2 (Os−S in Figure 2) and Cp_2Re_2 (Re−S in Figure 3) have the formal metal−metal bond orders of five and six required by the 18-electron rule. However, the lowest energy structure of Cp_2Re_2 ($Re-T_1$ in Figure 3) is a triplet with a formal Re−Re bond order of five giving the rhenium atoms the 17-electron configurations required for a binuclear triplet. The triplet structure of Cp_2Os_2 (Os−T in Figure 2) has the quintuple bond required to give both osmium atoms an 18-electron configuration. The triplet spin multiplicity in Os−T arises from the two unpaired electrons in the δ components of the Os−Os bond, which are only "half bonds" with single electrons.

The 18-electron rule does not apply to the dimetallocenes of tungsten and tantalum, since the required bond orders of seven and eight require more bonding orbitals than are available, after considering the three orbitals required for the

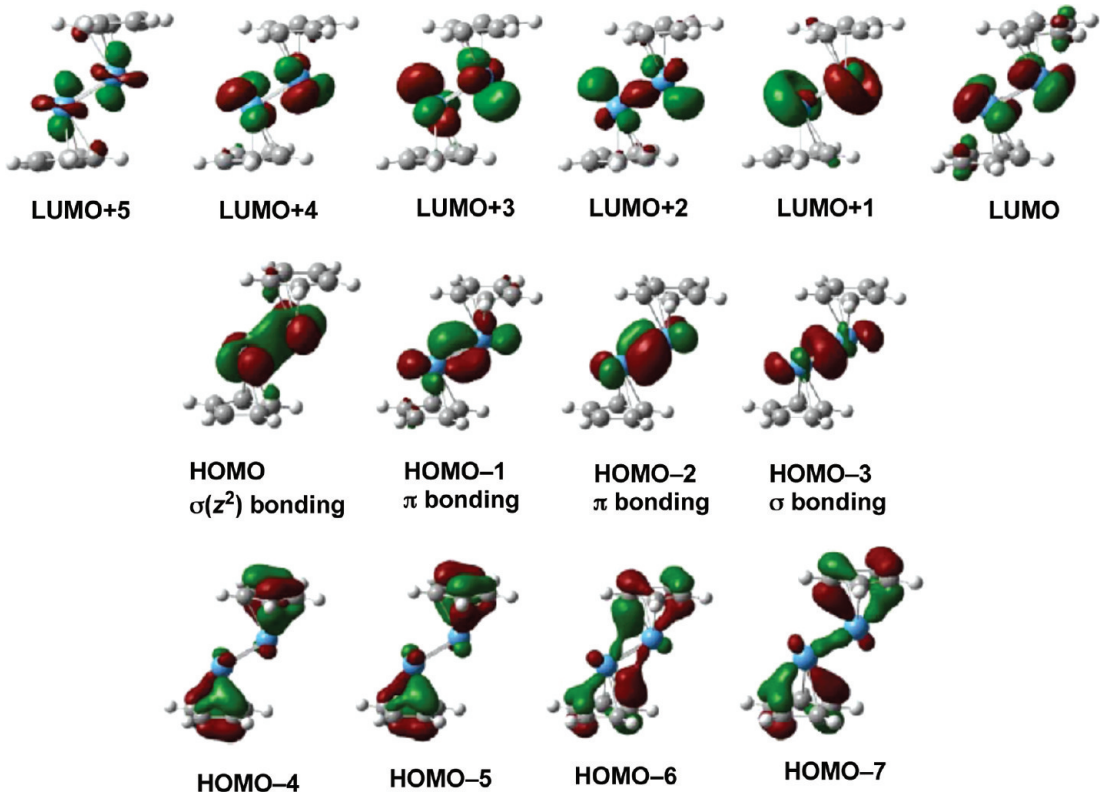


Figure 10. The frontier MOs for structure Ta−S_1 of Cp_2Ta_2. (a) Top: The LUMOs up to LUMO+5. (b) Middle: The HOMOs down to HOMO−3, relating to the Ta−Ta bonding. (c) Bottom: The next lower occupied MOs (HOMO−4 down to HOMO−7), relating largely to the metal−ring bonding.

metal−ring bonding. The W−W bond in singlet Cp_2W_2 (W−S in Figure 4) is a formal quintuple bond, giving both tungsten atoms 16-electron configurations. The Ta−Ta bond in singlet Cp_2Ta_2 is a formal quadruple bond giving both tantalum atoms the 14-electron configuration.

Metal−metal triple bonds are readily obtained in dimetallocene carbonyl chemistry, as indicated by the known stable compounds $Cp_2V_2(CO)_5$,[52,53] $Cp_2M_2(CO)_4$ (M = Cr,[54−56] Mo[57,58]), and $Cp_2M'_2(CO)_3$ (M′ = Mn,[59] Re[60]), all of which have formal M≡M triple bonds. Such M≡M triple bonds have one σ and two orthogonal π components, much like the C≡C triple bond in acetylene. In order to increase the formal metal−metal bond order above three, it is necessary to add either one or two δ components or a second σ component constructed by overlap of d(z^2) orbitals. The original experimentally achieved examples of metal−metal quadruple and quintuple bonds, namely $Re_2Cl_8^{2-}$ (ref 3) and ArylCrCrAryl (ref 4), respectively, supplement the $\sigma + 2\perp\pi$ triple bond with one or two full δ two-electron bonds, respectively. In this respect, the formal Os−Os quintuple bond in singlet Cp_2Os_2 (Os−S in Figure 2) is of the same type, i.e., $\sigma + 2\pi + 2\delta$, as the Cr−Cr quintuple bond in the Cr(I) aryl derivatives ArylCrCrAryl.[4]

The δ components of metal−metal multiple bonds are rather weak, as indicated by the chemistry of compounds with metal−metal quadruple bonds,[1] as well as the overlap in the relevant bonding molecular orbitals. For example, visual inspection of the two δ bonding orbitals HOMO−1 and HOMO−2 in singlet Cp_2Os_2 (Os−S in Figure 2) clearly indicates weaker overlap than the corresponding σ and π bonds (HOMO−4, HOMO−5, and HOMO−6). In the formal W−W quintuple bond of the electron poorer singlet ditungstocene W−S (Figures 4 and 9), there is only one δ component arising from the HOMO. The other component of this W−W quintuple bond is a second σ component from overlap of the d(z^2) orbitals (HOMO−1). In addition the formal Ta−Ta quadruple bond in singlet Ta−S (Figures 5 and 10) does not have any δ components but consists of two σ components and two π components.

None of the dimetallocenes discussed in this paper have yet been synthesized. However, such dimetallocenes are potentially accessible by the dehalogenation of $CpMX_n$ derivatives by reagents such as alkali metals. In order to stabilize these highly unsaturated Cp_2M_2 structures it might be necessary to introduce bulky substituents on the Cp rings in order to block further reactions of the metal−metal multiple bonds. Such strategies have been used to prepare unusual multiple bonds in main group element derivatives such as Ga≡Ga triple bonds in $[RGa{\equiv}GaR]^{2-}$ (R = bulky aryl ligand),[61] B═B double bonds in L→BH═BH←L,[62] and Si═Si double bonds in L→Si═Si←L (L = bulky carbene ligand).[63]

Acknowledgment. We are indebted to the 111 Project (B07012) and the National Natural Science Foundation (20873045) of China as well as the U.S. National Science Foundation (Grants CHE-0749868 and CHE-0716718) for support of this research.

Supporting Information Available: Tables S1−S5: Theoretical harmonic vibrational frequencies for Cp_2M_2 (M

= Os, Re, W, Ta) using the BP86 method. Tables S6−S24: Theoretical Cartesian coordinates Cp_2M_2 (M = Os, Re, W, Ta) using the MPW1PW91 method. Figures S1−S18: The frontier MOs for Cp_2M_2 (M = Os, Re, W, Ta) using the BP86 method. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Cotton, F. A.; Walton, R. A. *Multiple Bonds Between Metal Atoms*, 2nd Ed.; Clarendon: Oxford, 1993; pp 1−27.

(2) Radius, U.; Breher, F. *Angew. Chem., Int. Ed.* **2006**, *45*, 3006.

(3) Cotton, F. A.; Harris, C. B. *Inorg. Chem.* **1965**, *4*, 330–334.

(4) Nguyen, T.; Sutton, A. D.; Brynda, M.; Fettinger, J. C.; Long, G. J.; Power, P. P. *Science* **2005**, *310*, 844.

(5) Frenking, G. *Science* **2005**, *310*, 796.

(6) Radius, U.; Breher, F. *Angew. Chem., Int. Ed.* **2006**, *45*, 3006.

(7) Brynda, M.; Gagliardi, L.; Widmark, P.-O.; Power, P. P.; Roos, B. O. *Angew. Chem., Int. Ed.* **2006**, *45*, 3804.

(8) Roos, B. O.; Borin, A. C.; Gagliardi, L. *Angew. Chem., Int. Ed.* **2007**, *46*, 1469.

(9) Merino, G.; Donald, K. J.; D'Acchioli, J. S.; Hoffmann, R. *J. Am. Chem. Soc.* **2007**, *129*, 15295.

(10) Brynda, M.; Gagliardi, L.; Roos, B. O. *Chem. Phys. Lett.* **2009**, *471*, 1.

(11) Wagner, F. R.; Noor, A.; Kempe, R. *Nature Chem.* **2009**, *1*, 529.

(12) La Macchia, G.; Gagliardi, L.; Power, P. P.; Brynda, M. *J. Am. Chem. Soc.* **2008**, *130*, 5104.

(13) Wolf, R.; Ni, C.; Nguyen, T.; Brynda, M.; Long, G. J.; Sutton, A. D.; Fischer, R. C.; Fettinger, J. C.; Hellman, M.; Pu, L.; Power, P. P. *Inorg. Chem.* **2007**, *46*, 11277.

(14) Tsai, Y.-C.; Hsu, C.-W.; Yu, J.-S. K.; Lee, G.-H.; Wang, Y.; Kuo, T.-S. *Angew. Chem., Int. Ed.* **2008**, *47*, 7250.

(15) Hsu, C.-W.; Yu, J.-S. K.; Yen, C.-H.; Lee, G.-H.; Wang, Y.; Tsai, Y.-C. *Angew. Chem., Int. Ed.* **2008**, *47*, 9933.

(16) Noor, A.; Wagner, F. R.; Kempe, R. *Angew Chem. Int. Ed.* **2008**, *47*, 7246.

(17) Kreisel, K. A.; Yap, G. P. A.; Dmitrenko, O.; Landis, C. R.; Theopold, K. H. *J. Am. Chem. Soc.* **2007**, *129*, 14162.

(18) Resa, I.; Carmona, E.; Gutierrez-Puebla, E.; Monge, A. *Science* **2004**, *305*, 1136.

(19) Wang, Y.; Quillan, B.; Wei, P.; Wang, H.; Yang, X.-J.; Xie, Y.; King, R. B.; Schleyer, P. v. R.; Schaefer, H. F., III; Robinson, G. H. *J. Am. Chem. Soc.* **2005**, *127*, 11944.

(20) Xie, Y.; Schaefer, H. F., III; King, R. B. *J. Am. Chem. Soc.* **2005**, *127*, 2818.

(21) Ehlers, A. W.; Frenking, G. *J. Am. Chem. Soc.* **1994**, *116*, 1514.

(22) Delly, B.; Wrinn, M.; Lüthi, H. P. *J. Chem. Phys.* **1994**, *100*, 5785.

(23) Li, J.; Schreckenbach, G.; Ziegler, T. *J. Am. Chem. Soc.* **1995**, *117*, 486.

(24) Jonas, V.; Thiel, W. *J. Phys. Chem.* **1995**, *102*, 8474.

(25) Barckholtz, T. A.; Bursten, B. E. *J. Am. Chem. Soc.* **1998**, *120*, 1926.

(26) Niu, S.; Hall, M. B. *Chem. Rev.* **2000**, *100*, 353.

(27) Macchi, P.; Sironi, A. *Coord. Chem. Rev.* **2003**, *238*, 383.

(28) Buhl, M.; Kabrede, H. *J. Chem. Theory Comput.* **2006**, *2*, 1282.

(29) Tonner, R.; Heydenrych, G.; Frenking, G. *J. Am. Chem. Soc.* **2008**, *130*, 8952.

(30) Ziegler, T.; Autschbach, J. *Chem. Rev.* **2005**, *105*, 2695.

(31) Waller, M. P.; Bühl, M.; Geethanakshmi, K. R.; Wang, D.; Thiel, W. *Chem.—Eur. J.* **2007**, *13*, 4723.

(32) Hayes, P. G.; Beddie, C.; Hall, M. B.; Waterman, R.; Tilley, T. D. *J. Am. Chem. Soc.* **2006**, *128*, 428.

(33) Bühl, M.; Reimann, C.; Pantazis, D. A.; Bredow, T.; Neese, F. *J. Chem. Theory Comput.* **2008**, *4*, 1449.

(34) Besora, M.; Carreon-Macedo, J.-L.; Cowan, J.; George, M. W.; Harvey, J. N.; Portius, P.; Ronayne, K. L.; Sun, X.-Z.; Towrie, M. *J. Am. Chem. Soc.* **2009**, *131*, 3583.

(35) Ye, S.; Tuttle, T.; Bill, E.; Simkhorich, L.; Gross, Z.; Thiel, W.; Neese, F. *Chem.—Eur. J.* **2008**, *14*, 10839.

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

(37) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

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

(39) Zhao, Y.; Pu, J.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2004**, *6*, 673.

(40) Perdew, J. P. In *Electronic Structure of Solids*, 1991 ed.; Ziesche, P., Esching, H., Eds.; Akademic Verlag: Berlin, 1991; p 11.

(41) Zhao, S.; Wang, W.; Li, Z.; Liu, Z. P.; Fan, K.; Xie, Y.; Schaefer, H. F. *J. Chem. Phys.* **2006**, *124*, 184102.

(42) Feng, X.; Gu, J.; Xie, Y.; King, R. B.; Schaefer, H. F. *J. Chem. Theory Comput.* **2007**, *3*, 1580.

(43) Andrae, D.; Haussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor.Chim. Acta* **1990**, *77*, 123.

(44) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823.

(45) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.

(46) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(47) Zhang, J.; Huang, K.-W.; Szalda, D. J.; Bullock, R. M. *Organometallics* **2006**, *25*, 2209.

(48) Hoyano, J. K.; Graham, W. A. G. *Chem. Comm.* **1982**, 27.

(49) Alvarez, M. A.; Garcia, M. E.; Riera, V.; Ruiz, M. A.; Falvello, L. R.; Bois, C. *Organometallics* **1997**, *16*, 354.

(50) Cotton, F. A.; Frenz, B. A.; Etner, J. A.; Walton, R. A. *Inorg. Chem.* **1976**, *15*, 1630.

(51) Barden, C. J.; Rienstra-Kiracole, J. C.; Schaefer, H. F. *J. Chem. Phys.* **2000**, *113*, 609, and references cited therein.

(52) Cotton, F. A.; Kruczynski, L.; Frenz, B. A. *J. Organomet. Chem.* **1978**, *160*, 93.

(53) Huffman, J. C.; Lewis, L. N.; Caulton, K. G. *Inorg. Chem.* **1980**, *19*, 2755.

(54) Curtis, M. D.; Butler, W. M. *J. Organomet. Chem.* **1978**, *155*, 131.

(55) King, R. B.; Efraty, A.; Douglas, W. M. *J. Organomet. Chem.* **1973**, *60*, 125.

(56) Potenza, J.; Giordano, P.; Mastropaolo, D.; Efraty, A. *Inorg. Chem.* **1974**, *13*, 2540.

(57) King, R. B.; Bisnette, M. B. *J. Organomet. Chem.* **1967**, *8*, 287.

(58) Huang, J. S.; Dahl, L. F. *J. Organomet. Chem.* **1983**, *243*, 57.

(59) Herrmann, W. A.; Serrano, R.; Weichmann, J. *J. Organomet. Chem.* **1983**, *246*, C57.

(60) Hoyano, J. K.; Graham, W. A. G. *Chem. Comm.* **1982**, 27.

(61) Su, J.; Li, X.-W.; Crittendon, R. C.; Robinson, G. H. *J. Am. Chem. Soc.* **1997**, *119*, 5471.

(62) Wang, Y.; Quillian, B.; Wei, P.; Wannere, C. S.; Xie, Y.; King, R. B.; Schaefer, H. F.; Schleyer, P. R.; Robiinson, G. H. *J. Am. Chem. Soc.* **2007**, *129*, 12412.

(63) Wang, Y.; Xie, Y.; Wei, P.; King, R. B.; Schaefer, H. F., III; Schleyer, P. v. R.; Robinson, G. H. *Science* **2008**, *321*, 1069.

CT900564P

JCTC Journal of Chemical Theory and Computation

Calibration of Cholesky Auxiliary Basis Sets for Multiconfigurational Perturbation Theory Calculations of Excitation Energies

Jonas Boström,[†] Mickaël G. Delcey,[†] Francesco Aquilante,[‡] Luis Serrano-Andrés,[¶] Thomas Bondo Pedersen,[§] and Roland Lindh*,[†]

Department of Theoretical Chemistry, Chemical Center, University of Lund, P. O. Box 124 S-221 00 Lund, Sweden, Department of Physical Chemistry, Sciences II, University of Geneva, Quai E. Ansermet 30, 1211 Geneva 4, Switzerland, Instituto de Ciencia Molecular, Universitat de València, P. O. Box 22085, ES-46071 Valencia, Spain, and Centre for Theoretical and Computational Chemistry, Department of Chemistry, University of Oslo, P. O. Box 1033 Blindern, N-0315 Oslo, Norway

Received November 19, 2009

Abstract: The accuracy of auxiliary basis sets derived from Cholesky decomposition of two-electron integrals is assessed for excitation energies calculated at the state-average complete active space self-consistent field (CASSCF) and multiconfigurational second order perturbation theory (CASPT2) levels of theory using segmented as well as generally contracted atomic orbital basis sets. Based on 196 valence excitations in 26 organic molecules and 72 Rydberg excitations in 3 organic molecules, the results show that Cholesky auxiliary basis sets can be used without compromising the accuracy of the multiconfigurational methods. Specifically, with a decomposition threshold of 10^{-4} au, the mean error due to the Cholesky auxiliary basis set is 0.001 eV, or smaller, decreasing with increasing atomic orbital basis set quality.

1. Introduction

Given a Gaussian atomic orbital (AO) basis χ and an auxiliary basis φ, density fitting (DF) consists of determining a set of fitting coefficients C such that $\chi_\mu(\mathbf{r})\,\chi_\nu(\mathbf{r}) \approx \sum_K C^K_{\mu\nu}\,\varphi_K(\mathbf{r})$. In the most common DF approach, this amounts to the minimization of the following positive semidefinite error matrix

$$\Delta_{\mu\nu} = \left(\chi_\mu\chi_\nu - \sum_K C^K_{\mu\nu}\varphi_K|\chi_\mu\chi_\nu - \sum_K C^K_{\mu\nu}\varphi_K\right) \qquad (1)$$

where $(\cdot|\cdot)$ is a two-electron integral in Mulliken notation. Clearly, the error matrix can be made arbitrarily small, i.e., essentially exact, by a suitable choice of auxiliary basis set.

As an alternative to standard auxiliary basis sets, which are preoptimized to reproduce specific energy contributions (Coulomb, exchange, second-order dynamic correlation), we have recently proposed to determine the auxiliary basis functions on-the-fly by means of Cholesky decomposition (CD) of the entire or parts of the two-electron integral matrix in the AO basis.[1−4] An upshot of this approach is that the decomposition threshold directly specifies an upper bound to the error matrix $\Delta_{\mu\nu}$ in eq 1. In the full-CD approach,[5,6] the entire molecular two-electron integral matrix is Cholesky decomposed until all elements of the error matrix are below the specified decomposition threshold. One-center CD (1C-CD)[1] differs from full-CD in that only those elements of the error matrix for which χ_μ and χ_ν are centered on the same atom are bounded by the decomposition threshold. The atomic CD (aCD)[1] auxiliary basis sets are constructed by a decomposition of the atomic integral matrix, and like 1C-CD, only one-center errors are bounded by the decomposition threshold. Finally, the recently proposed atomic compact CD (acCD)[4] auxiliary basis sets are constructed from the aCD

* To whom correspondence should be addressed- E-mail: roland.lindh@teokem.lu.se.
[†] Lund University.
[‡] University of Geneva.
[¶] Universitat de València.
[§] University of Oslo.

10.1021/ct900612k
Published on Web 02/08/2010

ones by reducing the number of primitive Gaussians in the product functions. By removing the auxiliary functions with the highest angular momenta, the aCD and acCD sets can be brought to a size (number of auxiliary basis functions) similar to the standard sets, thus mitigating the issues raised concerning the efficiency of CD auxiliary basis sets.[7] The CD auxiliary basis sets offer several advantages such as complete freedom in selection of the atomic orbital (AO) basis set, an unbiased nature with respect to the wave function or density functional model, and nearly arbitrary control over the associated error. For a detailed discussion of the CD approach, we refer to the recent review by Pedersen et al.[3]

We have previously demonstrated the accuracy of the full-CD, 1C-CD, aCD, and acCD auxiliary basis sets for calculations of electronic ground state energies at the Hartree−Fock (HF), second-order Møller−Plesset (MP2), and hybrid as well as nonhybrid density functional theory (DFT) levels of theory with a range of AO basis sets.[8] Here, we proceed with the accuracy assessment of the CD auxiliary basis sets in conjunction with the complete active space self-consistent field (CASSCF)[9−11] and multiconfigurational second order perturbation theory (CASPT2)[12−14] approaches to the computation of vertical electronic excitation energies. Whereas the earlier investigation[8] monitored the accuracy of the CD auxiliary basis set in association with the ground states of the closed-shell molecules in the G2/97 test suite, the present investigation aims at measuring the accuracy of excited electronic states. For this purpose, the Schreiber test suite[15,16] has been employed. Recalling that even in the recent past AO basis sets beyond 300−400 functions were prohibitive, a number of recent applications employing 1000−1500 Gaussian basis functions[11,14,17−19] have clearly demonstrated the computational advantage of the CD-based CASSCF/CASPT2 approach. A major goal of the present study therefore is to establish accuracy standards, i.e., decomposition thresholds, for future reference.

Hättig and co-workers[20−23] have used auxiliary basis sets optimized for dynamic correlation (at the MP2 level) for the calculation of excitation energies and excited state properties with the second-order coupled cluster (CC2) model,[24] and Pedersen et al.[25−29] have used full-CD to calculate electric dipole polarizability and optical rotation from the CC2 linear response function which, formally, includes a summation over all (singlet) excited electronic states. A major difference between these two approaches is that in the CD case the same auxiliary basis set is used in all stages of the calculation including the computation of the HF reference function. That is, the same CD auxiliary basis set is used for Coulomb, exchange, and dynamic correlation contributions. At the DFT level, studies of the performance of preoptimized auxiliary basis sets for excitation energy calculations limited to the nonhybrid BP86 functional[30,31] have been reported by Bauernschmitt et al. and Rappoport and Furche.[32,33] Neese and co-workers[34,35] have extended these calibrations to a hybrid DFT functional. However, as we can judge from the technical details of the latter two studies, the RI approximation was not used in the ground state hybrid DFT calculations. Nevertheless, in these studies, the performance of auxiliary basis sets in accurately approximating the J and K type MO integrals used in the time dependent DFT (TDDFT) equations has been documented. To our knowledge, no benchmark studies have been presented in which a single auxiliary basis set has been calibrated with respect to both the ground state DFT energies and TD-DFT excitation energies using hybrid functionals. Considering the fact that the same CD auxiliary basis set is used for Coulomb, exchange, and static and dynamic correlation as modeled by the CASSCF/CASPT2 approach, there are strong reasons to believe that the results presented here can be directly transferred to nonhybrid DFT functionals and other wave function models used for excited state calculations.

2. Computational Details

The purpose of the present benchmark study is not to assess the accuracy of the CASPT2/CASSCF method in comparison with other quantum chemical models or experiments. Rather, we wish to test that the CD auxiliary basis sets can in an unbiased way be used to describe excited states, i.e., the ability of the CD auxiliary basis sets to accurately represent two-electron integrals involving virtual molecular orbitals. We use the CASSCF/CASPT2 protocol as a successful representative of the methods for computing excitation energies. Two test suites are employed. The first was designed by Schreiber et al.[15] and includes 196 vertical excitation energies, both singlet and triplet excited states, for 26 organic molecules. The excitations are mainly of a $\pi-\pi^*$ or $n-\pi^*$ nature with some $\sigma-\pi^*$ excitations. The second set was designed to test for Rydberg states. These calculations involve valence and Rydberg singlet and triplet states for ethene, *trans*-1,3-butadiene, and formamide. Here, we report statistics based on the 72 transitions involving the Rydberg states.

For the first series, the details of the computations are as follows: For molecules of high symmetry, the conventional as well as DF-based CASPT2 method, as currently implemented, fails to correctly preserve the degeneracy of states belonging to multidimensional irreducible representations. This would lead to reference data for degenerate states with a small artificial symmetry breaking. Even when this has no practical consequences for the interpretation of the excited states, complications in the analysis of the CD auxiliary basis set error are avoided by excluding the triazine and benzene molecules from the original test suite in this work. Geometries, active spaces, and other information needed to perform the calculations were chosen exactly as in the work of Schreiber et al.,[15] although we have extended the study to include four different AO basis sets. In addition to the original TZVP[36] basis set, the ANO-RCC-VXZP[37,38] (X = D, T, Q) AO basis sets were used. We note that, whereas the TZVP basis set was designed for the HF method, the ANO-RCC-VXZP sets were specifically designed in combination with the CASSCF/CASPT2 paradigm. However, the aim of this study is not to investigate the quality of the AO basis set with respect to the accuracy of excitation energies in comparison with experimental or "exact" excitation energies. Rather, we will only comment on the ability of

the CD auxiliary basis set to approximate integrals with the four different AO basis sets used in this study.

For the second series of benchmark calculations (the Rydberg benchmark), basis sets of the ANO-L type[37] contracted to C, N, O [4s3p1d]/H [2s1p] were employed (almost a full VTZP basis), supplemented by a set of 1s1p1d diffuse functions centered in the molecular cation charge centroid and whose exponents and coefficients were determined elsewhere.[39−41] The employed active spaces included the full valence π space (and the oxygen lone pair in formamide) plus the nine 3s3p3d Rydberg orbitals, leading to spaces (electrons,orbitals) (2,11), (4,13), and (6,13) for ethene, *trans*-1,3-butadiene, and formamide, respectively. The low-lying valence and Rydberg singlet and triplet states were computed, including the Rydberg series π (HOMO) $\rightarrow$ 3s3p3d and also $n \rightarrow$ 3s3p3d for formamide. The multistate (MS) CASPT2 method[42] was employed in order to take into account the valence-Rydberg mixing effect displayed at the CASSCF level of theory. The imaginary level-shift technique[43] was used with a parameter of 0.1 au to prevent the presence of weakly coupling intruder states. Geometries and basis sets were those used previously for ethene,[39] *trans*-1,3-butadiene,[39] and formamide.[40]

Calculations have been performed in these two series with the full-CD, aCD, and acCD auxiliary basis sets obtained with decomposition thresholds of 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6} au. The auxiliary basis set pruning technique (skipping higher angular components, SHAC), as used in our previous benchmark study,[8] was explored too. While the 1C-CD has been excluded from the present work for implementation reasons—the method has not been implemented for cases with point group symmetry—the aCD and acCD results should be representative of 1C-CD, as they were in our previous report on ground state energies.[8] The computed total and excitation energies were compared to reference energies compiled with conventional CASSCF/CASPT2.[13,44] For uracil, cytosine, thymine, adenine, and octatetraene with ANO-RCC-VQZP in the first test series, the numbers of AO basis functions are 560, 590, 675, 700, and 740, respectively. This number of functions is prohibitive for conventional CASSCF/CASPT2 calculations. In these cases, full-CD calculations with a Cholesky threshold of 10^{-10} au were used as a reference.

All calculations were performed with the MOLCAS 7 program package[45] using the CD-based implementations described by Aquilante et al.[11,14]

3. Results and Discussion

The accuracy of the CD auxiliary basis sets that we aim to establish should be compared to the accuracy of standard preoptimized auxiliary basis sets. It has been shown with TD-DFT calculations on 36 excited states for a set of small molecules (the largest being benzene) that errors in adiabatic excitation energies are not higher than 0.03 eV.[33] These results are in line with the earlier reported accuracy assessment by Bauernschmitt and co-workers on 19 vertical excitation energies.[32] We note that the results of Bauernschmitt et al.[32] and Rappoport and Furche[33] are based on the original TZVP auxiliary basis set of Eichkorn et al.,[46] which was designed for ground state calculations with nonhybrid DFT functionals. This auxiliary basis set was subsequently augmented by "downward extrapolation" including primitive Gaussians to improve the quality of the results. We stress two major differences between these benchmark studies and those of the current report. First, the present CD auxiliary basis sets are designed on-the-fly in a procedure with error control (in the sense of eq 1) via a single parameter, the decomposition threshold. Second, whereas the previous studies are based on nonhybrid DFT calculations, the current test includes Coulomb and exchange as well as correlation.

In the rest of this section, the accuracy of the CASSCF/CASPT2 excitation energies as a function of the CD approach and threshold, auxiliary basis set pruning, and AO basis set saturation is analyzed. We report mean errors, mean absolute errors, maximum errors, and standard deviations for excitation energies at the CASPT2 and CASSCF levels of approximation in Tables 1−8. The CASPT2 results are representative for the CASSCF results; hence, the CASSCF results, Tables 5−8, are included in the Supporting Information. Our findings and observations can be summarized as follows.

Cancellation of Errors. In one of the earliest investigations into the accuracy of CD auxiliary basis sets,[1] absolute and activation energies of 20 reactions were analyzed. In that study, we observed a favorable cancellation of errors with respect to the activation energies. Typically, the auxiliary basis set error was reduced by a factor of 2−4 in these cases. For the computation of vertical excitation energies, we would expect error cancellation to be at its optimum. In Figure 1, results obtained with the TZVP basis set and the full-CD, aCD, and acCD auxiliary basis sets are displayed.

A reduction of the mean absolute error for the same CD threshold of as much as 1 order of magnitude is observed going from total energies to excitation energies (compare the center and right panels of Figure 1). For total energies we see significant differences with respect to the full-CD, aCD, and acCD approximations as compared with the excitation energies. Here, we also note that the difference between the aCD and acCD auxiliary basis sets with and without auxiliary basis set pruning (SHAC) is significant. However, for the excitation energies, the discrepancy between the aCD and acCD auxiliary basis sets, with or without SHAC, is nearly completely removed. These trends hold true for any of the other AO basis sets in this study. Furthermore, we note that the mean absolute errors of the excitation energies are well below 0.01 eV already for the highest CD threshold of 10^{-3} au. We also note that, while full-CD shows an exponential decay of the mean absolute excitation energy error as a function of the CD threshold, the aCD and acCD auxiliary basis sets exhibit an error which is virtually constant and on the order of 0.001 eV. This suggests that the CD threshold can be increased by 1 order of magnitude when calculating excitation energies compared to total energy calculations.[8] We also note that auxiliary basis set pruning can be used in the computation of excitation energies, thus speeding up the calculations by reducing the number of auxiliary functions.

Table 1. Mean Errors in eV for CASPT2 Excitation Energies as a Function of AO Basis Set, CD Threshold (in au), and CD Auxiliary Basis Set (without and with pruning)

		mean errors for CASPT2 calculations				
					(SHAC)	
basis set	CD-thres.	full-CD	aCD	acCD	aCD	acCD
TZVP	10^{-3} au	-3.7×10^{-3}	-1.0×10^{-3}	-1.2×10^{-3}	3.4×10^{-3}	3.2×10^{-3}
	10^{-4} au	-7.6×10^{-4}	-5.9×10^{-4}	-5.9×10^{-4}	3.8×10^{-4}	4.3×10^{-4}
	10^{-5} au	-1.1×10^{-4}	-1.5×10^{-4}	-1.5×10^{-4}	5.4×10^{-4}	5.4×10^{-4}
	10^{-6} au	6.6×10^{-6}	-1.7×10^{-4}	-1.7×10^{-4}	5.5×10^{-4}	5.5×10^{-4}
ANO-RCC-VDZP	10^{-3} au	-2.6×10^{-3}	4.6×10^{-5}	2.5×10^{-5}	1.3×10^{-3}	1.3×10^{-3}
	10^{-4} au	7.9×10^{-6}	5.2×10^{-5}	2.6×10^{-5}	1.2×10^{-3}	1.2×10^{-3}
	10^{-5} au	-8.6×10^{-6}	5.4×10^{-5}	6.0×10^{-5}	9.7×10^{-4}	9.9×10^{-4}
	10^{-6} au	-2.3×10^{-6}	5.7×10^{-5}	5.4×10^{-5}	9.2×10^{-4}	9.3×10^{-4}
ANO-RCC-VTZP	10^{-3} au	3.9×10^{-3}	4.2×10^{-5}	1.2×10^{-5}	1.6×10^{-4}	1.4×10^{-4}
	10^{-4} au	1.2×10^{-4}	3.1×10^{-5}	-1.2×10^{-5}	1.4×10^{-4}	1.2×10^{-4}
	10^{-5} au	5.3×10^{-6}	8.3×10^{-6}	2.4×10^{-6}	1.2×10^{-4}	1.2×10^{-4}
	10^{-6} au	-5.3×10^{-6}	2.3×10^{-6}	1.5×10^{-6}	1.1×10^{-4}	1.1×10^{-4}
ANO-RCC-VQZP	10^{-3} au	-1.7×10^{-3}	9.4×10^{-6}	-1.4×10^{-6}	5.1×10^{-5}	-2.5×10^{-5}
	10^{-4} au	-5.2×10^{-4}	2.2×10^{-7}	-1.9×10^{-6}	1.8×10^{-5}	-2.8×10^{-5}
	10^{-5} au	-3.1×10^{-5}	9.7×10^{-8}	1.2×10^{-6}	1.7×10^{-5}	1.6×10^{-5}
	10^{-6} au	2.6×10^{-6}	2.3×10^{-7}	1.0×10^{-7}	1.5×10^{-5}	1.4×10^{-5}
ANO-L + Rydberg[a]	10^{-3} au	-6.6×10^{-4}	1.5×10^{-5}	7.2×10^{-6}	-1.6×10^{-4}	-4.6×10^{-5}
	10^{-4} au	-4.2×10^{-5}	1.5×10^{-5}	1.7×10^{-5}	-1.8×10^{-4}	-1.3×10^{-4}
	10^{-5} au	9.2×10^{-6}	1.4×10^{-5}	1.5×10^{-5}	-1.5×10^{-4}	-2.0×10^{-4}
	10^{-6} au	7.9×10^{-6}	2.1×10^{-5}	2.0×10^{-5}	-1.2×10^{-4}	-1.5×10^{-4}

[a] ANO-L C, N, O [4s3p1d]/H[2s1p] with explicit molecule-centered [1s1p1d] Rydberg functions.

Table 2. Standard Deviations in eV for CASPT2 Excitation Energies as a Function of AO Basis Set, CD Threshold (in au), and CD Auxiliary Basis Set (without and with pruning)

		standard deviations for CASPT2 calculations				
					(SHAC)	
basis set	CD-thres.	full-CD	aCD	acCD	aCD	acCD
TZVP	10^{-3} au	4.7×10^{-3}	1.1×10^{-3}	1.1×10^{-3}	7.0×10^{-3}	6.7×10^{-3}
	10^{-4} au	5.7×10^{-4}	4.1×10^{-4}	4.1×10^{-4}	7.8×10^{-4}	8.2×10^{-4}
	10^{-5} au	1.2×10^{-4}	2.3×10^{-4}	2.2×10^{-4}	6.4×10^{-4}	6.4×10^{-4}
	10^{-6} au	4.2×10^{-5}	2.2×10^{-4}	2.2×10^{-4}	6.4×10^{-4}	6.4×10^{-4}
ANO-RCC-VDZP	10^{-3} au	6.5×10^{-3}	5.9×10^{-4}	4.3×10^{-4}	1.3×10^{-3}	1.2×10^{-3}
	10^{-4} au	6.5×10^{-4}	6.3×10^{-4}	6.7×10^{-4}	1.3×10^{-3}	1.4×10^{-3}
	10^{-5} au	1.3×10^{-4}	6.3×10^{-4}	5.5×10^{-4}	1.3×10^{-3}	1.3×10^{-3}
	10^{-6} au	1.1×10^{-5}	6.5×10^{-4}	6.4×10^{-4}	1.3×10^{-3}	1.2×10^{-3}
ANO-RCC-VTZP	10^{-3} au	4.4×10^{-2}	5.0×10^{-4}	1.3×10^{-4}	5.0×10^{-4}	1.6×10^{-4}
	10^{-4} au	1.5×10^{-3}	3.6×10^{-4}	9.5×10^{-5}	3.6×10^{-4}	1.5×10^{-4}
	10^{-5} au	6.0×10^{-5}	1.0×10^{-4}	4.0×10^{-5}	1.3×10^{-4}	9.5×10^{-5}
	10^{-6} au	8.4×10^{-5}	2.2×10^{-5}	1.3×10^{-5}	8.0×10^{-5}	8.0×10^{-5}
ANO-RCC-VQZP	10^{-3} au	2.4×10^{-2}	1.3×10^{-4}	5.5×10^{-5}	4.0×10^{-4}	4.7×10^{-4}
	10^{-4} au	4.2×10^{-3}	5.1×10^{-6}	1.1×10^{-5}	2.5×10^{-5}	5.5×10^{-4}
	10^{-5} au	1.9×10^{-4}	3.0×10^{-6}	1.8×10^{-5}	2.4×10^{-5}	1.9×10^{-5}
	10^{-6} au	3.6×10^{-5}	3.7×10^{-6}	2.3×10^{-6}	1.9×10^{-5}	2.3×10^{-5}
ANO-L + Rydberg[a]	10^{-3} au	1.7×10^{-3}	8.9×10^{-5}	9.9×10^{-5}	8.9×10^{-4}	8.8×10^{-4}
	10^{-4} au	1.9×10^{-4}	9.0×10^{-5}	9.1×10^{-5}	8.3×10^{-4}	9.2×10^{-4}
	10^{-5} au	6.4×10^{-5}	9.0×10^{-5}	9.0×10^{-5}	5.7×10^{-4}	6.1×10^{-4}
	10^{-6} au	2.9×10^{-5}	8.7×10^{-5}	8.7×10^{-5}	5.7×10^{-4}	5.9×10^{-4}

[a] ANO-L C, N, O [4s3p1d]/H[2s1p] with explicit molecule-centered [1s1p1d] Rydberg functions.

AO Basis Set Convergence and CD Approximations. It has previously been observed[8] that the accuracy of the CD auxiliary basis sets for a given set of CD parameters (threshold and high angular momentum eliminations) improves with increased AO basis set quality. In Figure 2, this can be analyzed in the case of excitation energies for ANO-RCC-VXZP.

Here, we again note that, as the AO basis set is improved in the sequence X = D, T, Q, the overall accuracy is improved. While it is natural that this trend is less clear for the full-CD results, it is significant that it occurs for the aCD and acCD auxiliary basis sets. In particular, for the largest AO basis set, a rather loose threshold for aCD can be used without affecting the accuracy of the computed excitation energy. Furthermore, in comparing the different CD approaches, it is noted that the aCD and acCD with a CD threshold of 10^{-3} au can be used as a standard for all practical purposes. At this level of approximation, the errors due to the use of CD auxiliary basis sets can be ignored.

Pruned aCD and acCD Auxiliary Basis Sets. By construction, the aCD and acCD auxiliary basis sets include high angular momentum components. In an *ad hoc* pruning of the auxiliary basis sets, based on the reasoning that the

Table 3. Maximum Errors in eV for CASPT2 Excitation Energies as a Function of AO Basis Set, CD Threshold (in au), and CD Auxiliary Basis Set (without and with pruning)

		maximum errors for CASPT2 calculations				
					(SHAC)	
basis set	CD-thres.	full-CD	aCD	acCD	aCD	acCD
TZVP	10^{-3} au	-2.0×10^{-2}	-8.1×10^{-3}	-7.5×10^{-3}	3.0×10^{-2}	2.9×10^{-2}
	10^{-4} au	-2.2×10^{-3}	-2.1×10^{-3}	-2.1×10^{-3}	2.6×10^{-3}	3.0×10^{-3}
	10^{-5} au	7.5×10^{-4}	-1.4×10^{-3}	-1.4×10^{-3}	2.2×10^{-3}	2.2×10^{-3}
	10^{-6} au	3.2×10^{-4}	-1.4×10^{-3}	-1.4×10^{-3}	2.2×10^{-3}	2.2×10^{-3}
ANO-RCC-VDZP	10^{-3} au	-3.3×10^{-2}	-7.1×10^{-3}	-5.5×10^{-3}	-7.1×10^{-3}	-5.5×10^{-3}
	10^{-4} au	-6.0×10^{-3}	-8.2×10^{-3}	-7.9×10^{-3}	-8.2×10^{-3}	-8.4×10^{-3}
	10^{-5} au	-1.1×10^{-3}	-8.2×10^{-3}	-6.9×10^{-3}	-9.4×10^{-3}	-9.4×10^{-3}
	10^{-6} au	-9.0×10^{-5}	-8.7×10^{-3}	-8.5×10^{-3}	-8.7×10^{-3}	-8.5×10^{-3}
ANO-RCC-VTZP	10^{-3} au	5.9×10^{-1}	6.9×10^{-3}	1.6×10^{-3}	6.9×10^{-3}	1.7×10^{-3}
	10^{-4} au	1.6×10^{-2}	5.0×10^{-3}	-1.2×10^{-3}	5.0×10^{-3}	-1.2×10^{-3}
	10^{-5} au	5.6×10^{-4}	1.4×10^{-3}	5.4×10^{-4}	1.4×10^{-3}	5.4×10^{-4}
	10^{-6} au	-1.1×10^{-3}	3.0×10^{-4}	1.3×10^{-4}	3.3×10^{-4}	3.6×10^{-4}
ANO-RCC-VQZP	10^{-3} au	-3.1×10^{-1}	1.8×10^{-3}	-6.7×10^{-4}	5.5×10^{-3}	-6.4×10^{-3}
	10^{-4} au	-5.5×10^{-2}	7.0×10^{-5}	-1.2×10^{-4}	2.2×10^{-4}	-7.7×10^{-3}
	10^{-5} au	-1.9×10^{-3}	4.0×10^{-5}	2.5×10^{-4}	2.1×10^{-4}	-6.1×10^{-5}
	10^{-6} au	5.0×10^{-4}	4.8×10^{-5}	3.0×10^{-5}	-6.1×10^{-5}	-1.7×10^{-4}
ANO-L + Rydberg[a]	10^{-3} au	4.7×10^{-3}	-4.7×10^{-4}	-4.6×10^{-4}	-2.9×10^{-3}	-2.9×10^{-3}
	10^{-4} au	5.9×10^{-4}	-4.7×10^{-4}	-4.4×10^{-4}	-2.9×10^{-3}	-3.5×10^{-3}
	10^{-5} au	4.9×10^{-4}	-4.7×10^{-4}	-4.7×10^{-4}	-1.9×10^{-3}	-2.0×10^{-3}
	10^{-6} au	2.4×10^{-4}	-4.6×10^{-4}	-4.6×10^{-4}	-1.8×10^{-3}	-1.9×10^{-3}

[a] ANO-L C, N, O [4s3p1d]/H[2s1p] with explicit molecule-centered [1s1p1d] Rydberg functions.

Table 4. Absolute Mean Errors in eV for CASPT2 Excitation Energies as a Function of AO Basis Set, CD Threshold (in au), and CD Auxiliary Basis Set (without and with pruning)

		mean absolute error for CASPT2 calculations				
					(SHAC)	
basis set	CD-thres.	full-CD	aCD	acCD	aCD	acCD
TZVP	10^{-3} au	4.4×10^{-3}	1.1×10^{-3}	1.3×10^{-3}	6.3×10^{-3}	6.0×10^{-3}
	10^{-4} au	8.4×10^{-4}	6.1×10^{-4}	6.1×10^{-4}	7.2×10^{-4}	7.7×10^{-4}
	10^{-5} au	1.3×10^{-4}	2.0×10^{-4}	2.0×10^{-4}	7.0×10^{-4}	7.0×10^{-4}
	10^{-6} au	1.7×10^{-5}	2.0×10^{-4}	2.1×10^{-4}	7.0×10^{-4}	7.0×10^{-4}
ANO-RCC-VDZP	10^{-3} au	4.4×10^{-3}	1.8×10^{-4}	1.4×10^{-4}	1.5×10^{-3}	1.5×10^{-3}
	10^{-4} au	3.3×10^{-4}	1.9×10^{-4}	1.8×10^{-4}	1.4×10^{-3}	1.5×10^{-3}
	10^{-5} au	4.4×10^{-5}	1.9×10^{-4}	1.8×10^{-4}	1.3×10^{-3}	1.3×10^{-3}
	10^{-6} au	5.9×10^{-6}	1.8×10^{-4}	1.8×10^{-4}	1.2×10^{-3}	1.2×10^{-3}
ANO-RCC-VTZP	10^{-3} au	5.2×10^{-3}	4.5×10^{-5}	2.3×10^{-5}	1.7×10^{-4}	1.5×10^{-4}
	10^{-4} au	2.4×10^{-4}	3.5×10^{-5}	1.4×10^{-5}	1.6×10^{-4}	1.5×10^{-4}
	10^{-5} au	1.9×10^{-5}	9.6×10^{-6}	7.0×10^{-6}	1.3×10^{-4}	1.3×10^{-4}
	10^{-6} au	1.0×10^{-5}	2.9×10^{-6}	2.5×10^{-6}	1.2×10^{-4}	1.2×10^{-4}
ANO-RCC-VQZP	10^{-3} au	4.0×10^{-3}	9.8×10^{-6}	7.1×10^{-6}	5.5×10^{-5}	6.6×10^{-5}
	10^{-4} au	5.5×10^{-4}	6.6×10^{-7}	2.0×10^{-6}	2.2×10^{-5}	6.5×10^{-5}
	10^{-5} au	3.5×10^{-5}	4.4×10^{-7}	1.5×10^{-6}	2.1×10^{-5}	2.0×10^{-5}
	10^{-6} au	4.8×10^{-6}	5.1×10^{-7}	3.4×10^{-7}	1.9×10^{-5}	1.9×10^{-5}
ANO-L + Rydberg[a]	10^{-3} au	1.6×10^{-3}	5.7×10^{-5}	7.0×10^{-5}	6.5×10^{-4}	6.1×10^{-4}
	10^{-4} au	1.5×10^{-4}	5.9×10^{-5}	6.1×10^{-5}	6.0×10^{-4}	6.8×10^{-4}
	10^{-5} au	2.9×10^{-5}	5.9×10^{-5}	5.8×10^{-5}	4.3×10^{-4}	4.7×10^{-4}
	10^{-6} au	8.6×10^{-6}	5.8×10^{-5}	5.7×10^{-5}	4.4×10^{-4}	4.4×10^{-4}

[a] ANO-L C, N, O [4s3p1d]/H[2s1p] with explicit molecule-centered [1s1p1d] Rydberg functions.

high angular momentum components contribute insignificantly to the energy in most cases, the higher angular components of the auxiliary basis set are eliminated. This pruning technique (SHAC), originally suggested by Eichkorn et al.,[46] was explored in our previous benchmark study.[8] It was concluded that, although the pruning reduced the CD auxiliary basis set convergence toward an exact representation of the two-electron integrals, this was of no consequence in most quantum-chemical studies. While it was noted in the first benchmark study on total ground state energies[8] that this technique indeed reduces the accuracy, the current investigation on CASPT2 and CASSCF excitation energies exhibits a close to perfect cancellation of errors. This remarkable feature is demonstrated for all types of AO basis sets and for both valence and Rydberg excited states (see Figures 1−3). The conclusion is that, for excitation energies, acCD auxiliary basis sets with a CD threshold of 10^{-3} au introduce an error which is insignificant. With this procedure, standard deviations in the computed excitation energies are below 0.01 eV.

CASSCF vs CASPT2. It is well-known that the AO basis set convergence of the CASSCF method is significantly different from that of the CASPT2 method. While methods

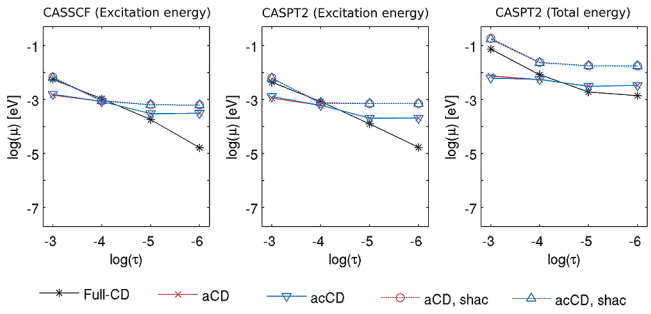


Figure 1. Mean absolute errors for excitation energies, μ, at the CASSCF (left panel) and CASPT2 (center panel) levels of theory, and mean absolute errors, μ, of the total energies at the CASPT2 level of theory (right panel) calculated with the TZVP basis set, with and without skipping of higher angular momenta (SHAC), as explained in section 3, plotted as a function of the CD threshold, τ.

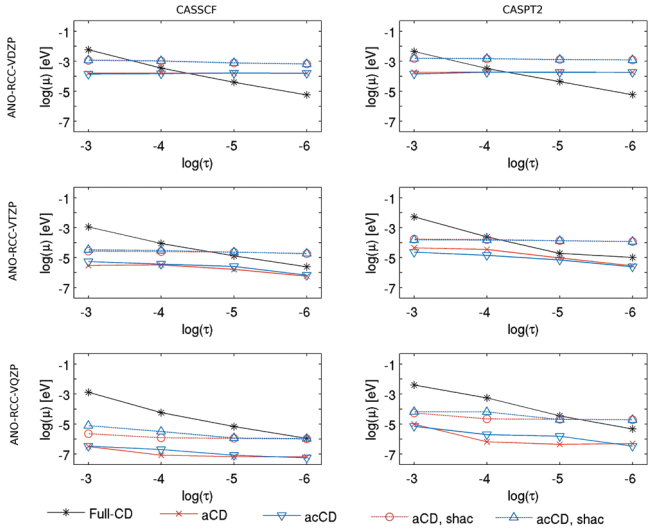


Figure 2. Mean absolute errors for excitation energies, μ, computed at the CASSCF (left) and CASPT2 (right) levels of theory with the ANO-RCC-VXZP, X = D, T, Q, basis sets, plotted as a function of the CD threshold, τ.

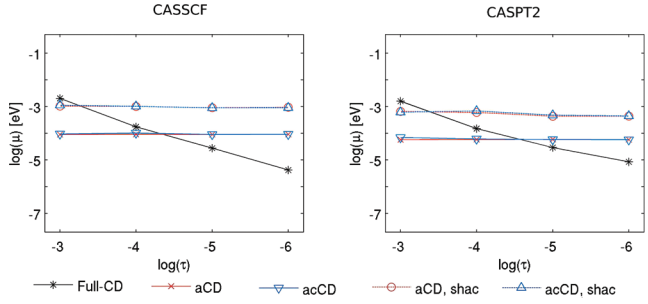


Figure 3. Mean absolute errors for excitation energies, μ, of Rydberg states at the CASSCF (left panel) and CASPT2 (right panel) levels of theory, with and without skipping of higher angular momenta (SHAC), plotted as a function of the CD threshold, τ.

like HF and CASSCF show close to AO basis set saturation already for triple-ζ quality AO basis sets, correlated methods like CASPT2 tend to need at least quadruple-ζ to achieve the same. Does this fact impact the requirements of the auxiliary basis sets? Looking at Figure 2 and comparing the left and right panel columns, we note hardly any significant accuracy difference in the computed CASSCF and CASPT2 excitation energies, with the possible exception that there is a slightly lower accuracy for the CASPT2 vs the CASSCF excitation energies for the ANO-RCC-VQZP basis set.

Rydberg vs Valence States. The accuracy assessments related to the 72 Rydberg states are presented in Figure 3.

In this study, we have employed explicit Rydberg basis sets[39−41] placed in the center of the molecule. This technique differs from that of adding atom-centered diffuse functions to the standard atomic AO basis sets, and its ability to accurately and conveniently reduce valence-Rydberg mixing and aid in identifying the Rydberg states has been documented previously.[41,42] We note that in the paper of Bauernschmitt et al.[32] the authors pointed out that the error in the computed Rydberg state excitation energies was larger, 0.08 eV, initially and that *ad hoc* decontraction or addition of diffuse auxiliary basis functions was required to achieve an accuracy similar to that of the errors found in the valence excitation energies, namely, 0.01 eV. In the full-CD, aCD, and acCD approaches, the explicit Rydberg AO basis does not require any special treatment as compared to any other AO basis set. Comparing Figure 3 with Figure 2 for the ANO-RCC-VTZP basis set, we note a slightly larger error in the case of the Rydberg excitations in combination with the aCD and acCD auxiliary basis sets. This is to some extent expected, considering the diffuse character of the explicit Rydberg basis and that these basis sets carry a lower significance in the aCD and acCD procedures. However, this can be completely ignored given that the mean absolute error in the excitation energies of the Rydberg states is on the order of 0.001 eV or better. For the full-CD approach, we find no significant error when comparing valence and Rydberg excitation energies. We conclude that, unlike the conventional DF auxiliary basis sets, no particular care has to be applied in the computation of Rydberg excitation energies for the present approaches.

Distribution of Errors. Finally, we analyze the distribution of errors and their maximum. A typical display of these is presented in Figure 4.

In particular, we note that the standard deviation in all cases, with the possible exception of full-CD with a CD threshold of 10^{-3} au, are below 0.01 eV. The same holds true for the maximum error. Furthermore, for thresholds tighter or equal to 10^{-4} au, standard deviations and maximum errors equal to or below 0.001 eV are observed. The accuracy here is certainly more than just in parity to assessments with external auxiliary basis sets[32,33]

4. Timings

Finally, we conclude this calibration paper with some brief notes on a typical representative case of an improvement in timings due to the use of the CD approximation for the CASPT2/CASSCF procedure. For this purpose, we have chosen to report the performances on some of the CASPT2 calculations described in ref 47. These systems are important intermediates from the reaction of O_2 with a Cu(I)−α-ketocarboxylate, and the accurate evaluation of the singlet−triplet splitting in each species is essential to the understand-

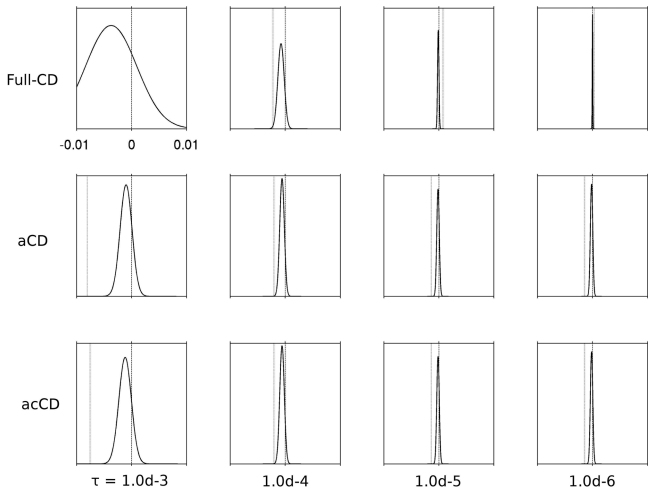


Figure 4. Display of the error distribution (in eV) as a function of the CD thresholds (in au) and DF approximations with the TZVP valence basis set. In this graph, the width of the Gaussian distribution is associated with the standard deviation. A dashed vertical line represents the maximum error found for the test set. The labeling of the upper left corner graph applies to all other graphs.

ing of the mechanism of activation of molecular oxygen by copper coordination complexes.[19] As described in detail in the original publication,[47] Cholesky-based CASSCF/CASPT2 calculations were perfomed with a decomposition threshold of 10^{-5} using ANO basis sets of double-ζ quality. This corresponds to a range of about 280 to 450 contracted Gaussian basis functions, depending on the system (point group symmetry not employed). In this range, the time spent to generate the DF-vectors is nearly independent of the particular choice of the Cholesky basis (full-CD, aCD, etc.), and so are the subsequent steps. For the smallest calculation, the CASPT2 step alone requires a wall-time of about 2 h using conventional two-electron integrals, compared to the 1 h of the full-CD (10^{-5}) implementation. The timings refer to an architecture of the type Intel(R) Xeon(TM) 3.20 GHz with 8 GB RAM and are those for the (8in8) choice of the active space (nearly identical timings result from the singlet or triplet calculation). Noticeably, the generation of the two-electron integrals in the MO basis shows alone a much better ratio: 85 vs 4885 s in wall-time (a factor 10 in CPU time). As discussed in the implementation paper,[14] the present DF-CASPT2 algorithm differs from the conventional only in the generation of the two-electron integrals in MO basis, whereas other computationally heavy tasks are left unchanged. In the above example, the task of solving the equations for the first-order wave function requires roughly 1 h of wall-time (45 min CPU time), and that explains why the resulting speedup is only a factor two. Moving toward the upper limit of 450 basis functions for our systems, the conventional calculations can hardly be afforded due to the large disk-space requirements. When possible, the DF-CASPT2 alone outperforms the conventional implementation by a factor of 5−8 in wall-time and with effectively no loss of accuracy (computed S−T splittings within 0.1 kcal/mol from conventional results). It should be also pointed out that the preliminary DF-CASSCF calculation can be performed at much lower costs than conventional calculations, and if included in the counting together with the integral/DF-vector generation, it gives rise to overall speedups that are much larger—not uncommonly 1−2 orders of magnitude. As an example, the generation of the Cholesky vectors for the smallest system requires only 5 min of wall-time, compared to 39 min needed to compute/store the AO two-electron integrals. The DF-CASSCF step is in this case about 4 times faster, 6 vs 25 min of wall-time.

5. Summary

In this study, we have reported the first accuracy assessments of the CD auxiliary basis set in association with the evaluation of vertical valence and Rydberg excitation energies computed with the CASSCF/CASPT2 protocol. These assessments clearly demonstrate the accuracy and flexibility of the CD auxiliary basis sets, specifically: (i) CD auxiliary basis sets offer excellent cancellation of errors. (ii) No significant differences were detected in comparing the errors associated with different AO basis sets. (iii) CD auxiliary basis set pruning can be employed safely. (iv) CD auxiliary basis sets give rise to essentially the same (insignificant) error in conjunction with CASSCF and CASPT2 excitation energy calculations. (v) The use of the CD procedure can reduce CASPT2/CASSCF wall-time by a factor of 4 up to 1−2 orders of magnitude. (vi) No special treatment of the CD auxiliary basis set is required in the computation of Rydberg excitation energies. (vii) The standard deviation observed by using the CD auxiliary basis sets in the computation of vertical excitation energies is well below 0.01 eV. (viii) CD threshold as high as 10^{-3} au can be used for calculating vertical valence and Rydberg excitation energies, giving mean and maximum errors in the range of 0.01 eV, and finally, (ix) for tighter thresholds, the CD auxiliary basis sets induce errors that are virtually completely insignificant.

Acknowledgment. The authors thank the Swedish Research Council (VR), the CoE Centre for Theoretical and Computational Chemistry (179568/V30), projects CTQ2007-61260 and CSD2007-0010 Consolider-Ingenio in Molecular Nanoscience of the Spanish MICINN/FEDER, and the Swedish Research Council (VR) through the Linnaeus Center of Excellence on Organizing Molecular Matter (OMM) for finacial support.

Supporting Information Available: Tables 5−8. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aquilante, F.; Lindh, R.; Pedersen, T. B. *J. Chem. Phys.* **2007**, *127*, 114107.

(2) Aquilante, F.; Lindh, R.; Pedersen, T. B. *J. Chem. Phys.* **2008**, *129*, 034106.

(3) Pedersen, T. B.; Aquilante, F.; Lindh, R. *Theor. Chem. Acc.* **2009**, *124*, 1–10.

(4) Aquilante, F.; Gagliardi, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2009**, *130*, 154107.

(5) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.

(6) Koch, H.; Sanchez de Meras, A.; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481–9484.

(7) Weigend, F.; Kattannek, M.; Ahlrichs, R. *J. Chem. Phys.* **2009**, *130*, 164106.

(8) Boström, J.; Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Theory Comput.* **2009**, *5*, 1545–1553.

(9) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157.

(10) Roos, B. O. *Int. J. Quantum Chem.* **1980**, *S14*, 175–189.

(11) Aquilante, F.; Pedersen, T. B.; Lindh, R.; Roos, B. O.; Sánchez de Merás, A.; Koch, H. *J. Chem. Phys.* **2008**, *129*, 024113.

(12) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

(13) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

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

(15) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(16) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.

(17) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2008**, *128*, 034104.

(18) Radon, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824–11832.

(19) Huber, S. M.; Ertem, M. Z.; Aquilante, F.; Gagliardi, L.; Tolman, W. B.; Cramer, C. J. *Chem.—Eur. J.* **2009**, *15*, 4886–4895.

(20) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(21) Hättig, C.; Köhn, A. *J. Chem. Phys.* **2002**, *117*, 6939–6951.

(22) Hättig, C.; Hald, K. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2111–2118.

(23) Kähn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *119*, 5021–5036.

(24) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

(25) Pedersen, T. B.; Sanchez de Meras, A. M. J.; Koch, H. *J. Chem. Phys.* **2004**, *120*, 8887–8897.

(26) Pedersen, T. B.; Koch, H.; Boman, L.; Sánchez de Merás, A. M. J. *Chem. Phys. Lett.* **2004**, *393*, 319–326.

(27) García Cuesta, I.; Pedersen, T. B.; Koch, H.; Sánchez de Merás, A. *Chem. Phys. Lett.* **2004**, *390*, 170–175.

(28) García Cuesta, I.; Sanchez Marín, J.; Pedersen, T. B.; Koch, H.; Sanchez de Meras, A. M. J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 361–365.

(29) Pedersen, T. B.; Kongsted, J.; Crawford, T. D.; Ruud, K. *J. Chem. Phys.* **2009**, *130*, 034310.

(30) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

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

(32) Bauernschmitt, R.; Häser, M.; Treutler, O.; Ahlrichs, R. *Chem. Phys. Lett.* **1997**, *264*, 573–578.

(33) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2005**, *122*, 064105.

(34) Neese, F.; Olbrich, G. *Chem. Phys. Lett.* **2002**, *362*, 170–178.

(35) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.

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

(37) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291–306.

(38) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.

(39) Serrano-Andrés, L.; Merchán, M.; Nebot-Gil, I.; Lindh, R.; Roos, B. O. *J. Chem. Phys.* **1993**, *98*, 3151–3162.

(40) Serrano-Andrés, L.; Fülscher, M. P. *J. Am. Chem. Soc.* **1996**, *118*, 12190–12199.

(41) Roos, B. O. The Complete Active Space Self-Consistent Field Method and its Applications in Electronic Structure Calculations. In *Advances in Chemical Physics; Ab Initio Methods in Quantum Chemistry - II*; Lawley, K. P., Ed.; John Wiley & Sons Ltd.: Chichester, England, 1987; Chapter 69, p 399.

(42) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299–306.

(43) Forsberg, N.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **1997**, *274*, 196.

(44) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.

(45) (a) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239. (b) Veryazov, V.; Widmark, P.-O.; Serrano-Andrés, L.; Lindh, R.; Roos, B. O. *Int. J. Quantum Chem.* **2004**, *100*, 626–653. (c) Aquilante, F.; De Vico, L.; Ferre, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrady, P.; Pedersen, T. B.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224–247.

(46) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(47) Huber, S. M.; Shahi, A. R. M.; Aquilante, F.; Cramer, C. J.; Gagliardi, L. *J. Chem. Theory Comput.* **2009**, *5*, 2967.

CT900612K

JCTC Journal of Chemical Theory and Computation

Characterization of Proton Coupled Electron Transfer in a Biomimetic Oxomanganese Complex: Evaluation of the DFT B3LYP Level of Theory

Ting Wang, Gary Brudvig, and Victor S. Batista*

Department of Chemistry, Yale University, PO Box 208107, New Haven, Connecticut 06520-8107

Received November 20, 2009

Abstract: The capabilities and limitations of the Becke-3-Lee−Yang−Parr (B3LYP) density functional theory (DFT) for modeling proton coupled electron transfer (PCET) in the mixed-valence oxomanganese complex $[(bpy)_2Mn^{III}(\mu\text{-}O)_2Mn^{IV}(bpy)_2]^{3+}$ (**1**; bpy = 2,2′-bipyridyl) are analyzed. Complex **1** serves as a prototypical synthetic model for studies of redox processes analogous to those responsible for water oxidation in the oxygen-evolving complex (OEC) of photosystem II (PSII). DFT B3LYP free energy calculations of redox potentials and pK_a's are obtained according to the thermodynamic cycle formalism applied in conjunction with a continuum solvation model. We find that the pK_a's of the oxo-ligands depend strongly on the oxidation states of the complex, changing by approximately 10 pH units (i.e., from pH ∼ 2 to pH ∼ 12) upon III,IV → III,III reduction of complex **1**. These computational results are consistent with the experimental pK_a's determined by solution magnetic susceptibility and near-IR spectroscopy as well as with the pH dependence of the redox potential reported by cyclic voltammogram measurements, suggesting that the III,IV → III,III reduction of complex **1** is coupled to protonation of the di-μ-oxo bridge as follows: $[(bpy)_2Mn^{III}(\mu\text{-}O)_2Mn^{IV}(bpy)_2]^{3+} + H^+ + e^- \rightarrow [(bpy)_2Mn^{III}(\mu\text{-}O)(\mu\text{-}OH)Mn^{III}(bpy)_2]^{3+}$. It is thus natural to expect that analogous redox processes might strongly modulate the pK_a's of oxo and hydroxo/water ligands in the OEC of PSII, leading to deprotonation of the OEC upon oxidation state transitions.

I. Introduction

Understanding the thermodynamics of proton coupled electron transfer (PCET) in oxomanganese complexes is essential for elucidating the mechanism of oxygen evolution by water oxidation, as catalyzed by the oxygen-evolving complex (OEC) of photosystem II (PSII).[1−7] The resulting insight on PCET is also necessary for the rational design of artificial photosynthetic systems.[8−11] This paper addresses the PCET mechanism in the mixed-valence oxomanganese dimer $[(bpy)_2Mn^{III}(\mu\text{-}O)_2Mn^{IV}(bpy)_2]^{3+}$ (**1**; bpy = 2,2′-bipyridyl), shown in Figure 1, as computationally characterized at the density functional theory (DFT) level with the Becke-3-Lee−Yang−Parr (B3LYP) hybrid density functional.[12,13]

Several oxomanganese complexes have been suggested as biomimetic models of the OEC of PSII,[8,14−19] including the mixed-valence oxomanganese dimer **1** originally synthesized by Nyholm and Turco[20] and characterized by X-ray crystallography by Plaksin et al.[21] In addition, Wang and Mayer[22] have studied an analogous complex with the 2,2′-bipyridyl ligand substituted by 1,10-phenanthroline. The cyclic voltammogram of **1** includes a reversible one-electron anodic couple at $E_{1/2}$ = 1.26 V (vs Ag/AgCl), assigned to the oxidation of the III,IV complex to the IV,IV state.[17,23] In addition, an irreversible one-electron cathodic wave with $E_{1/2}$ = 0.77 V at pH = 3.78 (Figure 1, top panel)[18] is thought to result from reduction of the mixed-valence III,IV complex to the III,III state. Furthermore, the Pourbaix diagram shows a linear dependence of $E_{1/2}$ with pH in the range pH = 3−9,

* Corresponding author. Fax: +1 203 432 6144, E-mail: victor.batista@yale.edu.

10.1021/ct900615b
Published on Web 01/29/2010

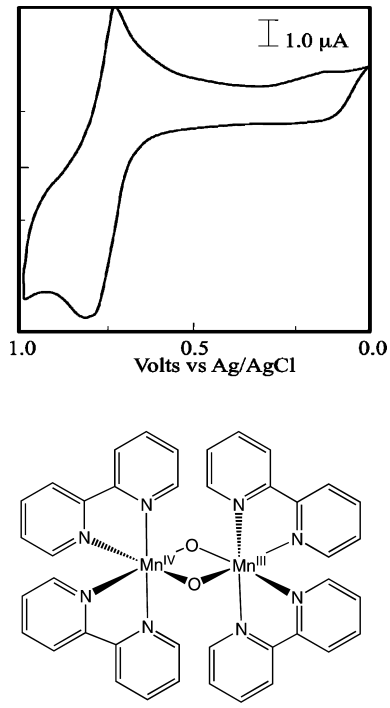


Figure 1. Cyclic voltammogram (top) for a 1 mM solution of complex **1** (bottom) in phosphate buffer at pH 3.78 as reported in ref 18.

with a ~59 mV/pH slope consistent with the one-electron one-proton couple:[18]

$$[(bpy)_2Mn^{III}(\mu\text{-}O)_2Mn^{IV}(bpy)_2]^{3+} + H^+ + e^- \rightarrow [(bpy)_2Mn^{III}(\mu\text{-}O)(\mu\text{-}OH)Mn^{III}(bpy)_2]^{3+} \quad (1)$$

The availability of electrochemical and spectroscopic data makes complex **1** ideally suited to investigate the capabilities and limitations of the DFT B3LYP level of theory as applied to studies of PCET in oxomanganese complexes. These studies complement our earlier work where we assessed the DFT B3LYP method as applied to the characterization of structural, electronic, and magnetic properties of synthetic oxomanganese complexes.[8] Our previous studies included Mn dimers, trimers, and tetramers where the metal centers are bridged by oxo ligands as well as models of the OEC of PSII analogous to the "3 + 1 Mn tetramer" model of the OEC of PSII.[8] Here, we extend these earlier studies to analyze the PCET reaction in complex **1**. Our investigations are based on free energy calculations of pK_a's and redox potentials, according to the thermodynamic cycle formalism in conjunction with a continuum solvation model.[24] Gas-phase free energies are first calculated, and then their values are corrected to account for solvation effects by using a dielectric continuum model. Such a standard computational procedure is one of the simplest approaches available to study redox and acid−base reactions in solution.[25−31]

Several studies have explored the capabilities of DFT methods for predictions of redox potentials of transition metal complexes.[24−41] Most of these earlier studies investigated functionals based on the generalized gradient approximation (GGA) such as BLYP,[42] BP86,[43] and PBE[44] as well as hybrid functionals (e.g., B3LYP[12,13]) originally developed and parametrized without including transition metal compounds in the reference data set. However, the description of redox potentials of oxomanganese complexes and the regulatory effect of oxidation state transitions on the pK_a's of oxo ligands bridging the Mn centers remain to be investigated. Exploring the capabilities and limitations of these methods is crucial to gaining insights on PCET mechanisms and to establishing the validity of currently available computational tools for the rational design of transition metal catalysts.

The paper is organized as follows. Section II outlines the computational methods applied for calculations of pK_a's and redox potentials. Section III presents our computational results and direct comparisons with experimental measurements. Concluding remarks and future research directions are outlined in section IV.

II. Computational Methods

All electronic structure calculations were carried out using the Jaguar suite of electronic structure programs.[45] Minimum energy configurations are obtained, as previously reported,[8] in broken symmetry (BS) states where the α and β electronic densities are localized on different metal centers. The B3LYP exchange-correlation functional with unrestricted Kohn−Sham wave functions (UB3LYP) yields ground state configurations for the reduced and oxidized forms of complex **1** with antiferromagnetically coupled high-spin manganese centers. Minimum energy configurations were obtained by using a mixed basis set, including the LACVP basis set to account for a nonrelativistic description of electron−core potentials (ECP's) for the Mn^{4+} and Mn^{3+} centers, the 6-31G (d) and 6-31G (2df) basis sets for bridging O^{2-} ions to include polarization functions for μ-oxo species, and the 6-31G basis sets for the rest of the atoms. All optimizations were followed by UB3LYP single point energy calculations based on Dunning's correlation-consistent triple-ζ basis set[46−48] cc-pVTZ(-f), including a double set of polarization functions. We have also tested the cc-pVTZ(-f)++ basis set, for which excellent agreement between calculated and experimental redox potentials was previously reported for other systems.[28] Both basis sets gave very comparable results.

Half-cell standard reduction potentials were obtained by computing the Gibbs free energy change $\Delta G_{red}(aq)$ due to the reduction of **1** in aqueous solution, as follows:

$$E^0 = -\frac{\Delta G_{red}(aq)}{nF} \quad (2)$$

with the Faraday constant $F = 23.06$ kcal mol^{-1} V^{-1} and $n = 1$ the number of electrons involved in the redox couple $[(bpy)_2Mn^{III}(\mu\text{-}O)_2Mn^{IV}(bpy)_2]^{3+}$/ $[(bpy)_2Mn^{III}(\mu\text{-}O)(\mu\text{-}OH)Mn^{III}(bpy)_2]^{3+}$. The values of $\Delta G_{red}(aq)$ were computed by using the half-reaction of the Born−Haber cycle, depicted in Figure 2, as follows:

$$\Delta G_{red}(aq) = \Delta G_{red}(g) + \Delta G_{sol}(III,III) - \Delta G_{sol}(III,IV) \quad (3)$$

Here, $\Delta G_{red}(g) = \Delta H_{red}(g) - T\Delta S_{red}(g)$ is the free energy change due to the reduction of **1** in the gas phase, with

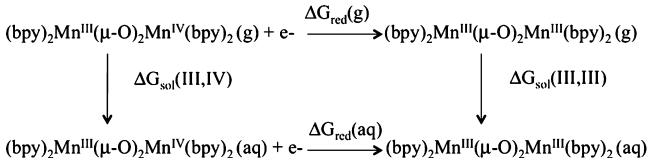


Figure 2. Born−Haber thermodynamic cycle for free energy calculations of redox potentials.

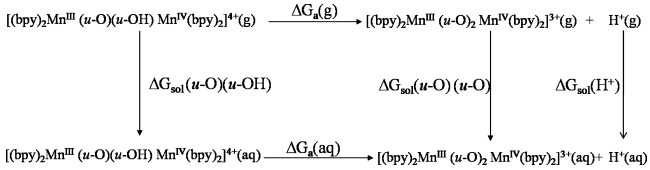


Figure 3. Born−Haber thermodynamic cycle used for free energy calculations of pK_a's.

$\Delta H_{red}(g) = \Delta H_{EA}(DFT) + \Delta H_{ZPE} + \Delta H_T$. $\Delta H_{EA}(DFT)$ is the electron attachment enthalpy, obtained at the DFT level for the complex in the gas phase, while the changes in the zero point energy ΔH_{ZPE} and corrections for molecular entropy changes $\Delta S_{red}(g)$ were based on vibrational frequency calculations. The solvation free energies of **1** in the oxidized and reduced forms, ΔG_{sol}(III,IV) and ΔG_{sol}(III,III), were computed by using the standard self-consistent reaction field (SCRF) approach,[49,50] based on accurate solutions of the Poisson−Boltzmann equation. Calculations were carried out for gas-phase geometries employing a dielectric constant of $\epsilon = 80.37$ (water) for the solvating continuum medium with a solvent radius of 1.40 Å. The effect of hydrogen bonding with solvent molecules or the coordination of buffer (phosphate) ligands to the metal centers is beyond the scope of this first study and will be addressed in a follow-up publication. Corrections due to changes in the thermal enthalpy ΔH_T were neglected.[28] All redox potentials are reported as relative potentials referenced to a silver chloride electrode (Ag/AgCl). The Ag/AgCl potential is 0.199 V more positive than that of the standard hydrogen electrode (SHE). Considering that the absolute potential of the SHE has been determined experimentally to be 4.43 eV,[51] we have subtracted 4.23 V from the absolute potentials to make direct comparisons to experimental data referenced to the Ag/AgCl.[18]

Our calculations of pK_a's were based on the following equation:

$$pK_a = \beta \Delta G_a(aq) \quad (4)$$

where $\beta = (k_B T)^{-1}$ corresponds to room temperature $T = 298.15$ K and k_B is the Boltzmann constant. The free energy change $\Delta G_a(aq)$ due to deprotonation of a μ-OH bridge for **1** in aqueous solutions was computed by using the half-reaction of the Born−Haber cycle, depicted in Figure 3, as follows:

$$\Delta G_a(aq) = \Delta G_a(g) + \Delta G_{sol}(\mu\text{-O},\mu\text{-O}) + \Delta G_{sol}(H^+) - \Delta G_{sol}(\mu\text{-O},\mu\text{-OH}) \quad (5)$$

where $\Delta G_a(g) = \Delta H_a(g) - T\Delta S_a(g)$ is the free energy change due to deprotonation in the gas phase, with an enthalpy change $\Delta H_a(g) = \Delta H_a(DFT) + \Delta H_{ZPE} + \Delta H_T$. Here, $\Delta H_a(DFT)$ is the energy change computed at the DFT level, due to deprotonation of the complex in the gas phase. The solvation free energies associated with the deprotonated and protonated forms of the complex, $\Delta G_{sol}(\mu$-O, μ-O) and $\Delta G_{sol}(\mu$-O, μ-OH), were also computed according to the SCRF approach,[49,50] as described above. We take the solvation free energy of a proton in water solvent to be $\Delta G_{sol}(H^+) = -260$ kcal mol^{-1}, as widely adopted in the literature.[51−53]

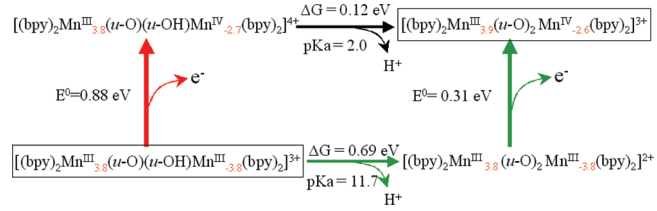


Figure 4. Thermodynamic energy diagram of PCET for complex **1** in aqueous solutions at pH = 0, as described by DFT B3LYP/cc-pVTZ(-f) free energy calculations of redox potentials and pK_a's based on the Born−Haber cycle method applied in conjunction with a continuum solvation model. Formal oxidation numbers are indicated as superscripts in Roman numbers, and the spin populations obtained according to the Mulliken population analysis are indicated as subscripts in red.

III. Results

Figure 4 shows the thermodynamic free energy diagram for PCET in complex **1** as obtained from DFT B3LYP/cc-pVTZ(-f) calculations. As described in section II, the redox potentials and pK_a's were obtained according to the Born−Haber cycle method, applied in conjunction with a continuum solvation model. Figure 4 shows that the reduced III,III state of complex **1** is expected to be protonated at pH < 11.7, with most of its population in the oxo-hydroxo form $[(bpy)_2Mn^{III}(\mu\text{-O})(\mu\text{-OH})Mn^{III}(bpy)_2]^{3+}$ (**1**$_{red}$), while **1** is protonated only at pH < 2. In addition, Figure 4 shows that the oxidation of **1**$_{red}$ is thermodynamically much easier when the complex is deprotonated in the $[(bpy)_2Mn^{III}(\mu\text{-O})_2Mn^{III}(bpy)_2]^{3+}$ state ($E^0 = 0.31$ V) than when it is protonated in the oxo-hydroxo state ($E^0 = 0.88$ V).

As shown in Figure 4, the protonated species **1**$_{red}$ can be oxidized via two possible pathways: (1) oxidation by a direct ionization process (red), requiring a rather high free energy of 0.88 eV, or (2) oxidation by a concerted removal of an electron from the complex and a proton from the μ-hydroxo bridge (green). The total energy requirement, thus, consists of two parts: 0.69 − 0.059*pH eV for the deprotonation step, and an extra 0.31 eV for subsequent oxidation of the deprotonated species. In an acidic environment, the oxidation related deprotonation steps are not spontaneous but driven by the externally applied electric field in a cyclic voltammogram (CV) experiment. Therefore, the CV peak position accounts for the free energy changes due to both deprotonation and oxidation. Because the electron ionization energy is constant, shifting of the potential for the redox reaction, thus, reflects the linear pH dependence of the associated deprotonation energy.

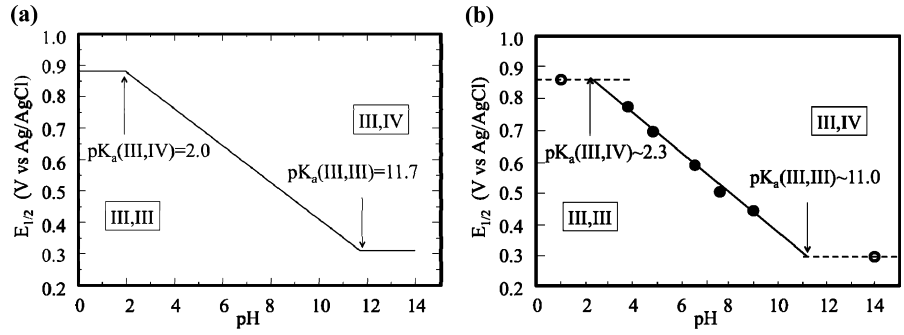


Figure 5. Pourbaix diagram for complex **1** in aqueous solutions, obtained (a) from free energy calculations of redox potentials at the DFT B3LYP/cc-pVTZ(-f) level of theory and (b) from experimental data (the circles) from ref 18.

The results reported in Figure 4 indicate that the oxidation of $\mathbf{1}_{red}$ to complex **1** is strongly coupled to deprotonation of the μ-OH bridge for a wide range of values of pH (i.e., pH = 2.0−11.7). For $\mathbf{1}_{red}$, the oxidation energy is constant at pH < 2.0 (oxidation takes the red path). It varies linearly at a rate of 59 mV/pH within the range of 2.0 < pH < 11.7 (green path with nonspontaneous deprotonation), as determined by the Nernst equation:

$$E_{1/2} = E^0 - \frac{0.05916}{n}\mathrm{pH} \qquad (6)$$

At pH > 11.7, the oxidation energy becomes a constant (the green path dominates with spontaneous deprotonation). Figure 5 shows the Pourbaix diagram, illustrating the pH dependence of $E_{1/2}$, as computed at the DFT B3LYP/cc-pVTZ(-f) level of theory and directly compared to experimental data.[18] The experimentally measured redox potentials for oxidation of **1** under different pH conditions are presented as circles in Figure 5b. The measured data points in the range of 4 < pH < 9 (the filled circles) exhibit linear pH dependence with a slope of ~59 mV/pH (the solid line). This particular value of the slope corresponds exactly with the slope expected for a system that loses one proton for each electron removed. Therefore, this linear relationship indicates a PCET in the range of 4 < pH < 9. The measured potentials at pH = 1 and pH = 14 (open circles) obviously deviate from the above linear relation. In such a strong acidic/basic solution environment, it is expected that the pH-independent (non-PCET) process dominates the oxidation of **1**, leading to the horizontal lines around pH = 1 and pH = 14 (dashed lines). Therefore, the crossover points of pH-independent and pH-dependent lines correspond to the pK_a's of the redox system.

The computational construction of the Pourbaix diagram (Figure 5a) is based on the ab initio calculation of pK_a's and redox potentials and predicts the redox potential of **1** for the entire pH range without relying on any kind of experimental data. According to Figure 4, the redox potential must be constant (0.88 V) at pH < 2.0 (oxidation takes the red path). It must vary linearly at a rate of 59 mV/pH within the range of 2.0 < pH < 11.7 (green path with nonspontaneous deprotonation), and it must be constant (0.31 V) at pH > 11.7 (the green path dominates with spontaneous deprotonation). The first crossover in Figure 5a corresponds to the conditions under which the red and green pathways, shown in Figure 4, are energetically identical. Therefore, the pH value of this crossover point provides the pK_a value of Mn(III, IV). Analogously, the pH value of the subsequent crossover point provides the pK_a value of Mn(III, III).

This comparison shows that there is a semiquantitative agreement between the calculated and experimental values of redox potentials throughout the whole range of pH.[54] The estimated errors are approximately ±1 unit of pH and ±60 mV for calculations of pK_a's and redox potential, respectively. These results, thus, suggest that the DFT B3LYP/cc-pVTZ(-f) level of theory could provide valuable descriptions of PCET processes in oxomanganese complexes, including other biomimetic catalysts and the OEC of PSII.

The molecular structure of the OEC of PSII has yet to be established.[3] Several structural models have been proposed, including the "3 + 1 Mn tetramer" with oxo-bridged high-valent Mn ions and Ca^{2+} found to be partially consistent with mechanistic studies of water oxidation and high-resolution spectroscopy.[2,7] However, it remains to be explored whether such a model is consistent with the well-known "redox leveling" effect by PCET. Such a regulatory mechanism is thought to avoid the buildup of charge in the cluster by deprotonation of water/hydroxo ligands, after oxidation state transitions, making all redox steps occur over a narrow range of potentials during the accumulation of 4 oxidizing equivalents. A similar redox leveling process is observed here for the oxomanganese complex **1**, for which the redox potential of the deprotonated state $[(bpy)_2Mn^{III}(\mu\text{-}O)_2Mn^{III}(bpy)_2]^{3+}$ (E^0 = 0.31 V) is significantly reduced as compared to the redox potential of the protonated state (E^0 = 0.88 V) manifesting such a redox leveling effect in good agreement with experimental data. The reported results thus partially validate the DFT B3LYP/cc-pVTZ(-f) level of theory for systems with common structural features, such as the OEC of PSII, when applied by combining the thermodynamic cycle formalism in conjunction with a continuum solvation model.

IV. Conclusions

We conclude that DFT B3LYP/cc-pVTZ(-f) calculations of redox potentials and pK_a's, obtained from standard gas-phase calculations and the subsequent correction for solvation effects by using a continuum solvation model, can provide valuable insights on the nature of PCET in oxomanganese complexes in aqueous environments. The reported computational results of redox potentials and pK_a's and the

favorable comparisons to experimental data from cyclic voltammogram measurements[18] and pK_a's determined by solution magnetic susceptibility and near-IR spectroscopy[17] demonstrate the capabilities of current DFT techniques as applied to modeling PCET in oxomanganese complexes. Both the regulatory effect of oxidation state transitions on the pK_a's of oxo ligands and the effect of deprotonation of hydroxo ligands on the redox potentials of metal centers can be properly modeled at the DFT B3LYP/cc-pVTZ(-f) level. Therefore, it is natural to expect that analogous redox-leveling processes could be modeled at the same level of theory for the OEC of PSII. Such calculations are currently underway in our group in an effort to establish structural and functional models of the OEC through rigorous comparisons to thermodynamic studies of water oxidation in PSII.

Acknowledgment. V.S.B. acknowledges supercomputer time from NERSC and financial support from the grant NIH 1R01-GM-084267-01. G.W.B. acknowledges support from the grant NIH GM32715.

Supporting Information Available: A description of the computational methods and the characterization of structural models in terms of the nuclear coordinates and computed thermodynamic data, including solvation energies, entropies, the spin population analysis, and the effect of oxidation state transition coupled to protonation/deprotonation events on the electrostatic potential atomic charges. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) McEvoy, J. P.; Brudvig, G. W. *Chem. Rev.* **2006**, *106*, 4455–4483.

(2) Sproviero, E. M.; Gascon, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *J. Am. Chem. Soc.* **2008**, *130*, 3428–3442.

(3) Sproviero, E. M.; Gascon, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *Coord. Chem. Rev.* **2008**, *252*, 395–415.

(4) Sproviero, E. M.; Gascon, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *J. Chem. Theory Comput.* **2006**, *2*, 1119–1134.

(5) Sproviero, E. M.; Gascon, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *Curr. Opin. Struct. Biol.* **2007**, *17*, 173–180.

(6) Ferreira, K. N.; Iverson, T. M.; Maghlaoui, K.; Barber, J.; Iwata, S. *Science* **2004**, *303*, 1831–1838.

(7) Sproviero, E. M.; Gascon, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *J. Am. Chem. Soc.* **2008**, *130*, 6728–6730.

(8) Sproviero, E. M.; Gascon, J. A.; McEvoy, J. P.; Brudvig, G. W.; Batista, V. S. *J. Inorg. Biochem.* **2006**, *100*, 786–800.

(9) Li, G. H.; Sproviero, E. M.; Snoeberger, R. C.; Iguchi, N.; Blakemore, J. D.; Crabtree, R. H.; Brudvig, G. W.; Batista, V. S. *Energy Environ. Sci.* **2009**, *2*, 230–238.

(10) McNamara, W. R.; Snoeberger, R. C.; Li, G.; Schleicher, J. M.; Cady, C. W.; Poyatos, M.; Schmuttenmaer, C. A.; Crabtree, R. H.; Brudvig, G. W.; Batista, V. S. *J. Am. Chem. Soc.* **2008**, *130*, 14329–14338.

(11) Abuabara, S. G.; Cady, C. W.; Baxter, J. B.; Schmuttenmaer, C. A.; Crabtree, R. H.; Brudvig, G. W.; Batista, V. S. *J. Phys. Chem. C* **2007**, *111*, 11982–11990.

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

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

(14) Cady, C. W.; Crabtree, R. H.; Brudvig, G. W. *Coord. Chem. Rev.* **2008**, *252*, 444–455.

(15) Mukhopadhyay, S.; Mandal, S. K.; Bhaduri, S.; Armstrong, W. H. *Chem. Rev.* **2004**, *104*, 3981–4026.

(16) Hagen, K. S.; Westmoreland, T. D.; Scott, M. J.; Armstrong, W. H. *J. Am. Chem. Soc.* **1989**, *111*, 1907–1909.

(17) Cooper, S. R.; Calvin, M. *J. Am. Chem. Soc.* **1977**, *99*, 6623–6630.

(18) Thorp, H. H.; Sarneski, J. E.; Brudvig, G. W.; Crabtree, R. H. *J. Am. Chem. Soc.* **1989**, *111*, 9249–9250.

(19) Sarneski, J. E.; Thorp, H. H.; Brudvig, G. W.; Crabtree, R. H.; Schulte, G. K. *J. Am. Chem. Soc.* **1990**, *112*, 7255–7260.

(20) Nyholm, R. S.; Turco, A. *Chem. Ind. (London)* **1960**, 74–75.

(21) Plaksin, P. M.; Palenik, G. J.; Stoufer, R. C.; Mathew, M. *J. Am. Chem. Soc.* **1972**, *94*, 2121–2122.

(22) Wang, K.; Mayer, J. M. *J. Am. Chem. Soc.* **1997**, *119*, 1470–1471.

(23) Morrison, M. M.; Sawyer, D. T. *J. Am. Chem. Soc.* **1977**, *99*, 257–258.

(24) Li, J.; Fisher, C. L.; Chen, J. L.; Bashford, D.; Noodleman, L. *Inorg. Chem.* **1996**, *35*, 4694–4702.

(25) Roy, L. E.; Batista, E. R.; Hay, P. J. *Inorg. Chem.* **2008**, *47*, 9228–9237.

(26) Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R. *J. Phys. Chem. A* **2009**, *113*, 6745–6750.

(27) Wang, T.; Friesner, R. A. *J. Phys. Chem. C* **2009**, *113*, 2553–2561.

(28) Baik, M. H.; Friesner, R. A. *J. Phys. Chem. A* **2002**, *106*, 7407–7412.

(29) Moens, J.; Geerlings, P.; Roos, G. *Chem.—Eur. J.* **2007**, *13*, 8174–8184.

(30) Moens, J.; Jaque, P.; De Proft, F.; Geerlings, P. *J. Phys. Chem. A* **2008**, *112*, 6023–6031.

(31) Moens, J.; Roos, G.; Jaque, P.; Proft, F.; Geerlings, P. *Chem.—Eur. J.* **2007**, *13*, 9331–9343.

(32) Tsai, M. K.; Rochford, J.; Polyansky, D. E.; Wada, T.; Tanaka, K.; Fujita, E.; Muckerman, J. T. *Inorg. Chem.* **2009**, *48*, 4372–4383.

(33) Cheng, T. Y.; Szalda, D. J.; Hanson, J. C.; Muckerman, J. T.; Bullock, R. M. *Organometallics* **2008**, *27*, 3785–3795.

(34) Muckerman, J. T.; Fujita, E.; Hoff, C. D.; Kubas, G. J. *J. Phys. Chem. B* **2007**, *111*, 6815–6821.

(35) Fujita, E.; Brunschwig, B. S.; Creutz, C.; Muckerman, J. T.; Sutin, N.; Szalda, D.; van Eldik, R. *Inorg. Chem.* **2006**, *45*, 1595–1603.

(36) Hou, H.; Muckerman, J. T.; Liu, P.; Rodriguez, J. A. *J. Phys. Chem. A* **2003**, *107*, 9344–9356.

(37) Uudsemaa, M.; Tamm, T. *J. Phys. Chem. A* **2003**, *107*, 9997–10003.

(38) Yang, X.; Baik, M. H. *J. Am. Chem. Soc.* **2006**, *128*, 7476–7485.

(39) Ayala, R.; Sprik, M. *J. Chem. Theory Comput.* **2006**, *2*, 1403–1415.

(40) Galstyan, A.; Knapp, E. W. *J. Comput. Chem.* **2009**, *30*, 203–211.

(41) De Groot, M. T.; Koper, M. T. M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1023–1031.

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

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

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

(45) *Jaguar*, 5.0 ed.; Schrodinger, Inc: Portland, OR, 2000.

(46) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(47) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(48) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

(49) Rashin, A. A.; Honig, B. *J. Phys. Chem.* **1985**, *89*, 5588–5593.

(50) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775–11788.

(51) Reiss, H.; Heller, A. *J. Phys. Chem.* **1985**, *89*, 4207–4213.

(52) Jang, Y. H.; Sowers, L. C.; Cagin, T.; Goddard, W. A. *J. Phys. Chem. A* **2001**, *105*, 274–280.

(53) Lim, C.; Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 5610–5620.

(54) The value of pK_a may change by several pH units, if the buffer coordination is considered, see experiments by Meyer et al in ref 22.

CT900615B

JCTC Journal of Chemical Theory and Computation

Protein Backbone Dynamics Simulations Using Coarse-Grained Bonded Potentials and Simplified Hydrogen Bonds

Tap Ha-Duong*

Laboratoire Analyse et Modélisation pour la Biologie et l'Environnement Université d'Evry-Val-d'Essonne Rue du Pere André Jarlan, 91025 Evry Cedex, France

Received August 5, 2009

Abstract: A new set of bonded potentials is introduced to model the flexibility of coarse-grained polypeptide chains. Based on a statistical analysis of known structures, the bonded potentials are sequence-dependent, and the secondary-structure propensity of each amino acid is partially reflected in the S_i—B_i—B_{i+1}—B_{i+2} pseudotorsion angle, where S_i and B_i denote the side-chain and backbone beads, respectively. To stabilize the secondary structures during simulations, the bonded force field must be balanced by a simplified model of the protein hydrogen bonds, based on dipole−dipole interactions. Tested on eight polypeptides with sequence lengths ranging from 17 to 98, using 200-ns molecular dynamics simulations, the coarse-grained model yields trajectories with RMSDs ranging from 3 to 8 Å from the experimental conformations. The less-structured regions of the simulated proteins exhibit the largest-amplitude movements.

1. Introduction

It is widely accepted that the conformational dynamics of backbone proteins plays an important role in their biological functions.[1] To mention just one example among many, the opening and reclosing motion of the two flaps that protect the active site of the HIV-1 protease is one of the key steps of its enzymatic mechanism.[2,3] The allosteric effect in proteins, which regulates their biological activities, is also well-known to involve conformational rearrangements of their backbones[4] and/or alterations of their dynamic properties.[5] These backbone motions can be probed by experimental techniques, particularly NMR spectroscopy, which can measure the angular mobility of the N—H bonds.[6] Protein structural changes can also be examined by theoretical methods, such as classical molecular dynamics simulations.[7] Nevertheless, both NMR and computational approaches generally meet difficulties in studying the broad-amplitude and long-time-scale movements of large proteins. Hence, the development of novel methodologies to investigate the functional internal motions of large biomolecular systems is still a very active research field.

Coarse-grained (CG) models of polymers,[8] particularly proteins, are now very popular, as the reduction in the number of particles enhances the exploration of phase space,[9] accelerates computer calculations, and provides insight into biological processes occurring on up to a microsecond time scale.[10,11] Among such simplified models, those at the residue level, which describe each amino acid with one or a few beads, succeed in combining computational efficiency with realistic descriptions of protein structural details. Thus, since the pioneer work of Levitt in 1976,[12] a large number of CG protein force fields have been developed, mainly to tackle the protein folding issue, but also to simulate the conformational dynamics of large proteins.[13,14] CG protein models have also been applied to the protein−protein docking problem, as bead softness can implicitly account for side-chain local flexibility and improve the predictions of matching interfaces between quasirigid proteins.[15]

The reliability of reduced protein models depends on a fine balance between the various force-field terms, and their interpretative strength relies on a simple formulation and a clear separation of the different physical driving forces. As in classical atomic models, CG protein force fields usually have a nonbonded (or long-range) contribution, which includes van der Waals and electrostatic interactions, and a

* E-mail: thaduong@univ-evry.fr.

10.1021/ct900408s
Published on Web 02/04/2010

bonded (or short-range) contribution, which determines the local geometry and flexibility of the polypeptide chains.[12,16−24] Whereas a physical basis can guide the building of nonbonded potentials between protein coarse grains,[25−29] the empirical parametrization of the bonded terms is not straightforward, as their ability to reproduce protein secondary and tertiary structures depends on the details of the nonbonded interactions, particularly the hydrogen bonds.

The flexibility of CG proteins can be efficiently modeled using elastic network models, which replace all of the interactions between pairs of beads that are separated by a distance lower than a cutoff parameter with quadratic potentials. Despite their simplicity and ease of implementation in molecular modeling programs, these one-parameter models can capture the essential features of the functional low-frequency large deformations of proteins.[30−34] One drawback of most elastic network models, however, is the absence of any explicit reference to the sequence of proteins, which hinders the study of the influence of mutations on protein dynamic behaviors. In addition, these models cannot describe large anharmonic motions and possible binding-induced structural changes of the polypeptide chains. To study these latter phenomena, one can hardly avoid the development of bonded potentials.

Effective bonded force fields for CG flexible proteins can be divided into two families that differ in terms of the level of resolution. In the high-level group, the peptide backbone is generally described with three united atoms, one for the nitrogen and its hydrogen atom, another for the α-carbon and its hydrogen atom, and the third for the carbonyl carbon and its oxygen atom.[9,22,23,35,36] In these descriptions, the backbone dihedral angles are similarly defined as in atomic models and can be directly calibrated to reproduce the Ramachandran energy landscapes. In addition, the two backbone united atoms NH and CO allow for the natural introduction of the hydrogen-bond interactions that stabilize secondary structures.[22,35,37] In the low-resolution models, the amino-acid backbone is represented with a single bead. In that case, the residue propensity to form secondary structures has to be implicitly encoded in the backbone pseudobending and/or pseudotorsion potentials,[18,24,31,38] and the hydrogen-bond stabilizing effect has to be introduced through empirical potentials[3,17,39] or electrostatic interactions.[16,38] These one-bead backbone models significantly reduce the number of local minima in the conformational space and generate overall less-frustrated energy landscapes.[9] However, in addition to the difficulty of calibrating the balance between the various energetic terms, most of these models require more or less preliminary information about native secondary or tertiary structures to simulate the protein dynamic conformation, such as in refs 3 and 24. Except for a few studies including those carried out by Scheraga, Liwo, and co-workers[40,41] and the recent one by Majek and Elber,[38] the conformational stability and dynamics of CG polypeptidic chains over long trajectories has seldom been examined using off-lattice unbiased one-bead backbone models.

This article presents an effort to build a general empirical bonded force field of CG proteins that allows their conformational changes to be studied by means of molecular dynamics (MD) simulations. This model of polypeptide flexibility is the natural continuation of the CG nonbonded potential that was recently derived from an all-atom force field by Basdevant et al.[29] The CG bonded potentials are completed with a simplified model of hydrogen bonds, formulated in terms of dipolar interactions and not biased toward any particular protein conformation or secondary structure. The CG protein model does not yet include a consistent description of solvation, especially to account for hydrophobic effects. Using instead a distance-dependent dielectric function as a crude model of hydration, the aim of this work is to bring out a minimal set of CG physical potentials that can reproduce the dynamic stability of proteins with MD simulations. This study therefore primarily focuses on the equilibrium structural properties of CG proteins and compares them with experimental observations, principally those provided by NMR spectroscopy.

2. Methods

2.1. Description of the Amino Acids. In the CG protein model presented herein, each amino acid is described with one bead for the backbone atoms and one or two beads for the side chain, depending on its size (Figure 1).[29] Each bead is located at the geometric center of the heavy atoms that it represents. As in classical all-atom force fields, it is assumed that the nonbonded contribution can be separated and expressed as sums of pairwise nonpolar and Coulombic energy functions depending on the grain-to-grain distances. In the present study, all of the beads are neutral, except for those corresponding to the two terminal backbone residues ($Nter^+$ and $Cter^-$) and the extremities of the charged side chains (ARG^+, ASP^-, GLU^-, and LYS^+). In the following discussion, the backbone grains are denoted B_i, and the side-chain grains are denoted S_i and S_i^*. The coarse-grained nonpolar potentials were obtained by numerical integration of the mean Lennard-Jones forces between pairs of amino acids. To calculate these latter quantities, 20 all-atom MD simulations for all possible homologue pairs of amino acids were performed in vacuo, using fully flexible molecules and a vanishing charge in order to capture the purely nonelectrostatic interaction. Then, to obtain a numerically tractable expression for the nonpolar energies, we fit all of the computed potentials of mean force with a unique mathematical function. The function that was found to best fit all 29 potentials, regardless of the size and softness of the coarse grains, consists of a repulsive part in r^{-6} and a Gaussian attractive part

$$V_{\mathrm{vdW}}(r_{ij}) = \varepsilon_{ij}\left\{\left(\frac{\lambda_{ij}}{r_{ij}}\right)^6 - \exp\left[-\left(\frac{r_{ij}}{\sigma_{ij}}\right)^2\right]\right\} \qquad (1)$$

The values of the parameters ε_{ii}, λ_{ii}, and σ_{ii} for the self-van der Waals interactions, as well as comparisons between CG energies and averaged all-atom Lennard-Jones potentials, are given in ref 29. To include cross-interactions and keep the number of parameters as small as possible, the empirical Lorentz−Berthelot mixing rules were applied: $\varepsilon_{ij} = (\varepsilon_{ii}\varepsilon_{jj})^{1/2}$, $\lambda_{ij} = (\lambda_{ii} + \lambda_{jj})/2$, and $\sigma_{ij} = (\sigma_{ii} + \sigma_{jj})/2$.

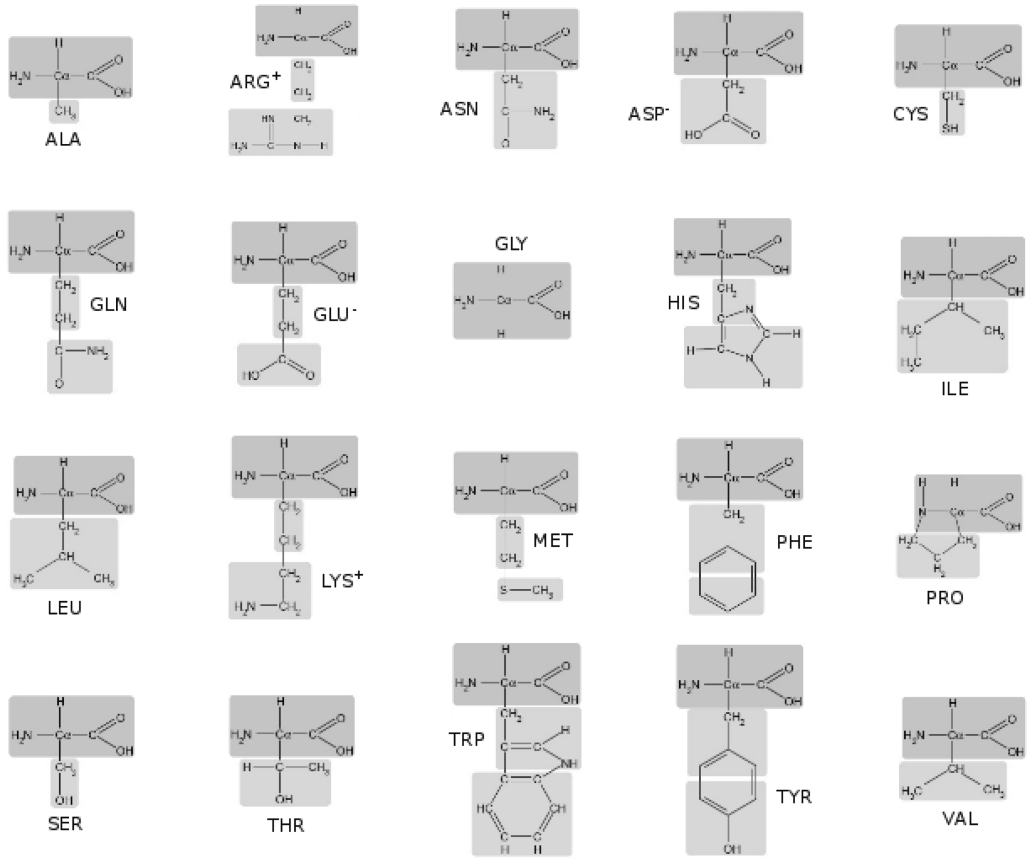


Figure 1. Presentation of the coarse-grained amino acids. Each gray rectangle represents a bead, whose position is located at the geometric center of the heavy atoms that form the coarse grain.

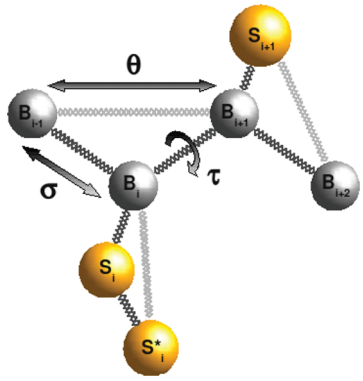


Figure 2. Schematic representation of the CG polypeptide model with its pseudobonds, pseudobends, and pseudotorsions. The backbone beads are denoted B_i, and the side chain beads are denoted S_i and S_i^*. Dark gray springs describe pseudobonds between the coarse grains, whereas light gray springs represent distance-dependent potentials used to account for the bending conformations.

2.2. Determination of the Bonded Potentials. The bonded potentials presented here are subdivided into three contributions (Figure 2), namely, a bond term depending on the length σ of all of the pseudobonds, a bending term expressed as a function of the distance θ between grains separated by two successive bonds, and a torsion term for all dihedral angles: $V_{\text{bonded}} = \sum V_{\text{bon}}(\sigma) + \sum V_{\text{ben}}(\theta) + \sum V_{\text{tor}}(\tau)$. To estimate these energy functions, a statistical analysis was first performed on a nonredundant set of 550 experimental structures of proteins (listed in Table 1 of the Supporting Information). Then, the knowledge-based potentials were extracted from the normalized probability distribution functions P, using Boltzmann inversion procedures: $V = -\gamma k_B T \ln(P)$, where k_B is the Boltzmann constant and T is the temperature. This approach, introduced by Miyazawa and Jernigan,[42] has been employed by many groups mainly to build short-range and long-range energy potentials that enable the discrimination of protein native folds from decoy structures.[20,43−46] In this study, this method was used to extract only the profiles of empirical bonded energy functions that govern the local conformation of the CG protein backbone, as the nonbonded contributions were previously determined from an all-atom protein model.[29] The empirical factors γ are weight scaling parameters introduced to balance the bonded energies against the nonbonded potentials that were determined using another approach. Preliminary trials on small peptides showed that, to generate stable MD trajectories, the bonded potentials needed to be stiffened using γ factors higher than 1. The tests revealed that values of $\gamma_{\text{bon}} = 4$, $\gamma_{\text{ben}} = 8$, and $\gamma_{\text{tor}} = 6$ for the bond, bending, and torsion potentials, respectively, can yield satisfactory results. Finally, the bonded potentials were fitted with tractable mathematical expressions to be implemented in an MD algorithm.

The database of 550 proteins includes both bound and unbound structures from the Protein Data Bank,[47] having less than 50% sequence similarity, refined to a crystallographic resolution lower than 3 Å, with no missing or unresolved heavy atoms. This set represents about 158000

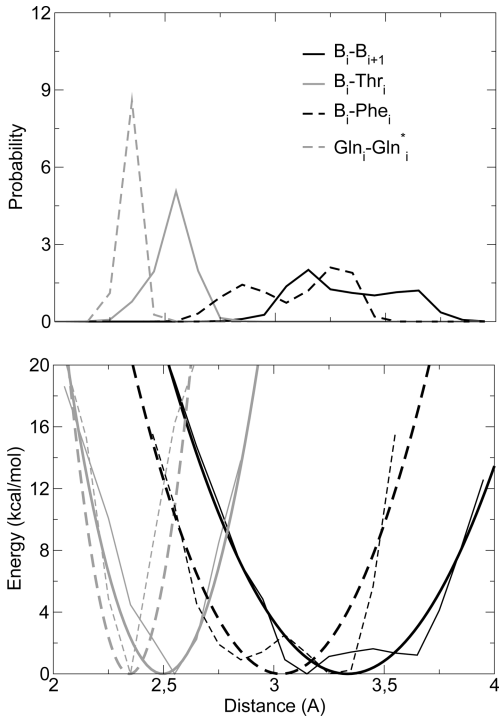


Figure 3. Top: Probability distribution functions for various pseudobond types. Bottom: Associated potentials (thin lines) and fitting energy functions (thick lines).

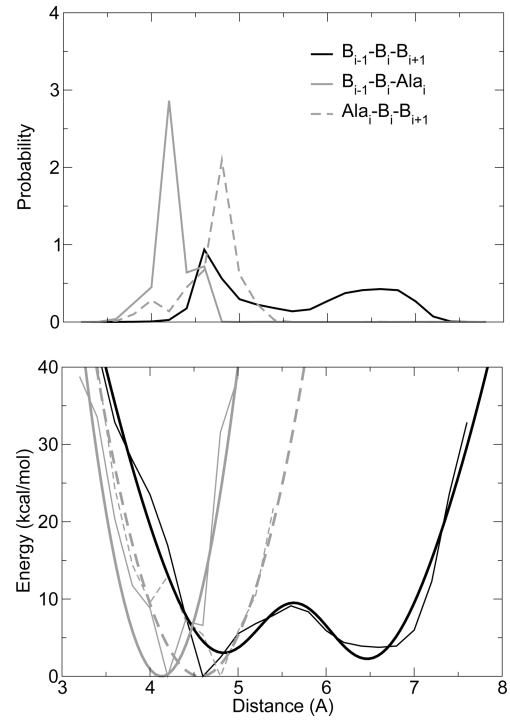


Figure 4. Top: Probability distribution functions for various pseudobend types. Bottom: Associated potentials (thin lines) and fitting energy functions (thick lines).

amino acids, of which 37% are considered in α-helices, 34% in β-strands, and 29% in nonstructured coils.

2.3. Bond, Bending, and Torsion Energy Functions. Figure 3 shows the probability distribution functions for the B_i—B_{i+1} and three others bond types. Most of the bond probability distributions have a peak shape around a single equilibrium value, whereas a few of them present a bimodal profile, including some virtual bonds B_i—S_i involving large side chains (Phe, Trp, Tyr) and especially the B_i—B_{i+1} bond distribution whose two peaks are associated with the helical and extended conformations of the backbone. Nevertheless, the length values between the two maxima are also significantly populated, and Boltzmann inversion yields a bond potential whose the second minimum is not very pronounced. Therefore, a single-well quadratic function was assumed to be a reasonable approximation of the B_i—B_{i+1} energy, as well as of all the other bond potentials (Figure 3)

$$V_{\text{bon}}(\sigma) = K_0(\sigma - \sigma_0)^2 \tag{2}$$

The energy functions associated with the bending angles were replaced with Urey−Bradley-like potentials depending on the distance between beads separated by two successive bonds (Figure 2). Preliminary works using bending-angle-dependent probability distributions generated a B_{i-1}—B_i—B_{i+1} energy profile whose second minimum associated with the β-structures was not very well-defined. For this reason, distance-dependent potentials are used instead of the angle-dependent ones, in order to better account for the two preferential B_{i-1}—B_i—B_{i+1} bending conformations. Because the side chains have a nonsymmetrical orientation relative to the direction of the protein backbone, it is important to differentiate between the B_{i-1}—B_i—S_i and S_i—B_i—B_{i+1} bending types, as illustrated in Figure 4. The bending probability distribution functions generally have a single-peak profile, except for eight bends that have a clear bimodal shape (B_{i-1}—B_i—B_{i+1}, B_{i-1}—B_i—His_i, B_{i-1}—B_i—Phe_i, B_{i-1}—B_i—Trp_i, B_{i-1}—B_i—Tyr_i, Asp_i—B_i—B_{i+1}, B_i—Phe_i—Phe^*_i, and B_i—Tyr_i—Tyr^*_i). Therefore, all of the bending potentials were fitted with quadratic functions, except the previously mentioned ones, for which double-well functions better fit the potentials (Figure 4). These functions were built by adding a harmonic term and a Gaussian term

$$V_{\text{ben}}(\theta) = K_1(\theta - \theta_1)^2 + K_2 \exp\left[-\left(\frac{\theta - \theta_2}{\theta_3}\right)^2\right] + K_3 \tag{3}$$

It should be noted that the two minima of the B_{i-1}—B_i—B_{i+1} generic energy function (Figure 4) are characteristic of α-helix and β-coil secondary structures, and that the tendency of the B_{i-1}—B_i—B_{i+1} triplet to adopt an α- or β-coil local conformation is indirectly modulated by the two other bends B_{i-1}—B_i—S_i and S_i—B_i—B_{i+1}. Furthermore, the bimodal profiles of the B_{i-1}—B_i—B_{i+1} probability distribution and potential were found by both Levitt[12] and Scheraga and co-workers[48,49] to be correlated with those of the adjacent backbone torsions. Nevertheless, it appears quite difficult to estimate how strong this correlation is and the extent to which it is a causal correlation rather than being due to a third physical factor, such as hydrogen bonding or other interactions. Thus, it is not clear whether this correlation has to be explicitly included in the CG bonded force field. The first version of the CG model presented herein neglects this correlation and assumes that the backbone internal coordinates are all independent degrees of freedom. Neglect of this

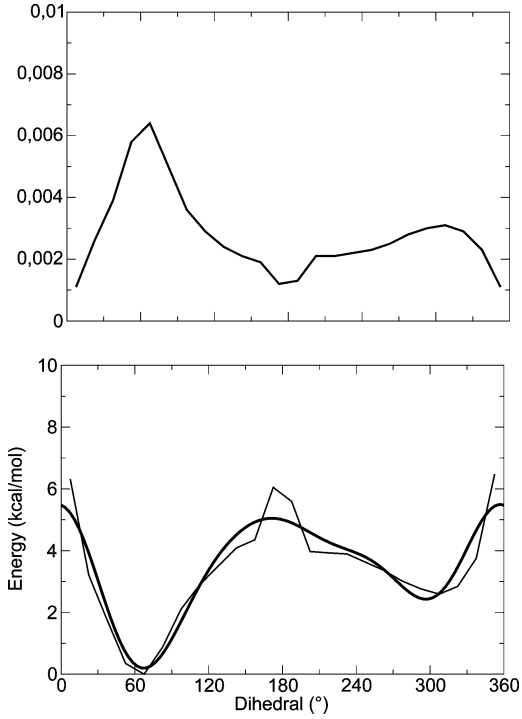


Figure 5. Top: Probability distribution function for the B_{i-1}—B_i—B_{i+1}—B_{i+2} pseudotorsion type. Bottom: Associated potential (thin line) and fitting energy function (thick line).

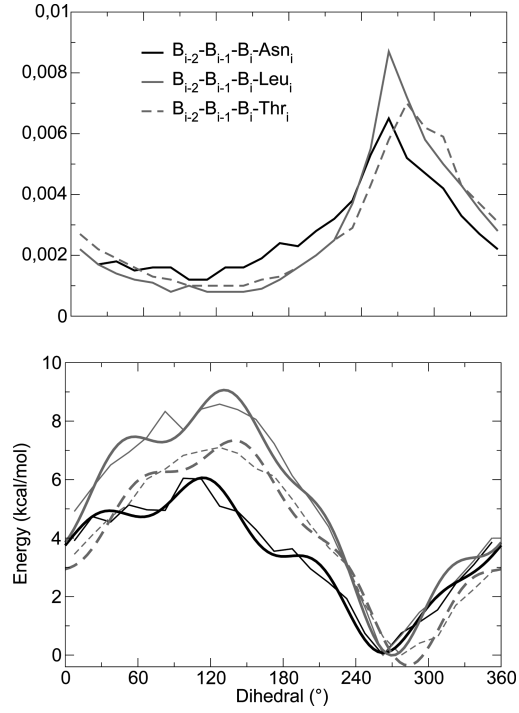


Figure 6. Top: Probability distribution functions for three B_{i-2}—B_{i-1}—B_i—S_i torsion types. Bottom: Associated potentials (thin lines) and fitting energy functions (thick lines).

correlation might have led to the need to stiffen the CG bonded potentials using the weight scaling parameters γ introduced in the previous section to generate stable MD trajectories. At present, further developments are being conducted to study the infuence of an explicit implementation of this correlation on the performance of the force field.

The pseudotorsion potentials are critical in determining a close description of the protein backbone secondary-structure propensity and possible large-amplitude dynamics. In the proposed model, the conformation of each backbone dihedral angle is locally determined by three potential types: the B_{i-1}—B_i—B_{i+1}—B_{i+2} torsion, which is sequence-independent, and the two distinct torsions B_{i-2}—B_{i-1}—B_i—S_i and S_i—B_i—B_{i+1}—B_{i+2}, which involve the side-chain grains. This approach differs significantly from other CG pseudotorsion models, which introduce only one potential for the B_{i-1}—B_i—B_{i+1}—B_{i+2} dihedral angle but whose the shape and parameters depend on the nature of the amino acids S_i and S_{i+1}.[16,18,20,38] The presented model allows for the consideration of only $1 + (2 \times 19)$, rather than 20×20, energy functions.

The torsion B_{i-1}—B_i—B_{i+1}—B_{i+2} probability distribution function and its associated potential (Figure 5) have general shapes similar to those found in several previous studies,[18,20] with two preferential conformations. However, it should be emphasized that, whereas the first minimum (around 70°) is clearly associated with α-helical structures, the second one (around 300°) predominantly corresponds to unstructured coils and widely overlaps the β-conformations, which have backbone torsion mean values around 190°. Hence, by itself, the B_{i-1}—B_i—B_{i+1}—B_{i+2} potential cannot yield stable β-strand structures, and these latter need to be stabilized by other local interactions. Compared to previous similar studies,[18,20] the difference in the position of the second minimum of the B_{i-1}—B_i—B_{i+1}—B_{i+2} potential arises slightly from the location of the B_i grains, which is at the geometric center of the backbone groups and not the Cα atoms. [9] It also certainly comes from differences in the number and proportion of α/β-coils in the protein databases that were used to extract the potentials. This is probably a significant drawback of knowledge-based approaches, but despite this limitation, they allow empirical energy functions to be generated easily and provide instructive insights into protein local short-range interactions.

For all of the amino acids except proline, the B_{i-2}—B_{i-1}—B_i—S_i probability and energy functions have similar profiles with a single preferential conformation between 240° and 300° (Figure 6). It should be noticed here that all of these torsion energies allow for the maintenance of the local chirality of the backbone beads B_i (with the exclusion of the Gly, N*ter*, and C*ter* residues). Indeed, for a given B_{i-2}—B_{i-1}—B_i—B_{i+1} backbone dihedral conformation, the position of the side chain S_i relative to the three beads B_{i-1}, B_i, and B_{i+1} is energetically determined by the single minimum of the torsion potential B_{i-2}—B_{i-1}—B_i—S_i (in addition to the bond length B_i—S_i and the two bending potentials B_{i-1}—B_i—S_i and S_i—B_i—B_{i+1}). In contrast, the S_i—B_i—B_{i+1}—B_{i+2} torsion types present more various behaviors that clearly depend on the nature of the amino acid. Their torsion potentials generally have two minima, one between 60° and 90° associated with β-strands and coils and a second around 210° that is characteristic of α-helices (Figure 7). One can distinguish three tendencies among these energy functions: For amino acids such as Ala, Arg, Glu, Gln, Lys, and Met, the minimum associated with the β-strand and coil conformations is clearly less deep than the α-related minimum. On the contrary, for residues such as Asn, Asp,

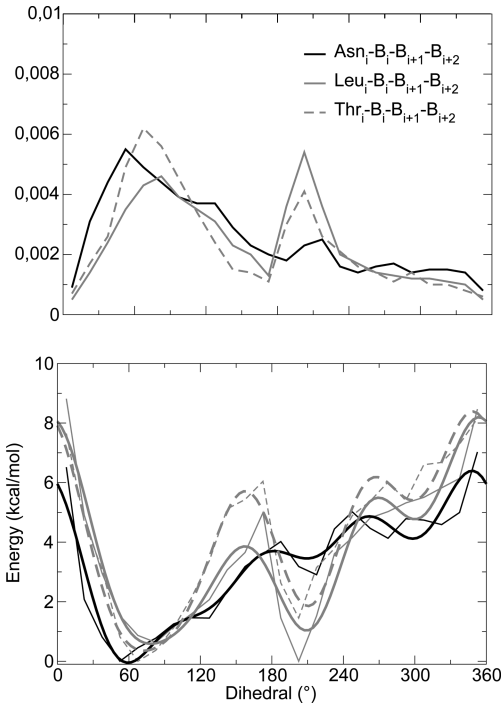


Figure 7. Top: Probability distribution functions for three S_i—B_i—B_{i+1}—B_{i+2} torsion types. Bottom: Associated potentials (thin lines) and fitting energy functions (thick lines).

Cys, Pro, Ser, and Thr, the well associated with the α structures is unfavored relative to the other one. In the third group, including the His, Ile, Leu, Phe, Trp, Tyr, and Val amino acids, the two minima coexist to a similar extent. Although these tendencies can hardly be quantitatively correlated to existing secondary-structure propensity scales, these observations reflect the overall residue probability to form secondary motifs previously computed by Chou and Fasman.[50] The chosen mathematical function to fit all torsion potentials is a fourth-order polynomial of sines, which is a good compromise between simplicity of the analytical form and accuracy of the fit

$$V_{\text{tor}}(\tau) = \sum_{n=1}^{4} A_n \sin^n(\tau - \tau_n) \qquad (4)$$

All of the parameters for the pseudobond, bending, and torsion potentials are given in Tables 2−4 of the Supporting Information.

2.4. Coarse-Grained Model of Hydrogen Bonds. The hydrogen bonds within proteins are particularly known to stabilize the protein secondary structures. These interactions, which are mainly electrostatic in nature, occur not only between amino acids separated by four or few residues, as in α-helices, but also between amino acids that could be very distant in the sequence, as in β-sheets. Following the idea of Liwo et al.,[16] a simplified model of protein hydrogen bonds can be introduced through dipole−dipole interactions between all pairs of backbone grains. Nevertheless, instead of placing a dipolar vector at the center of each backbone bead and calculating orientation-dependent interactions, a positively charged extra site H_i is attached to each grain B_i (which now carries an negative charge) in order to reproduce

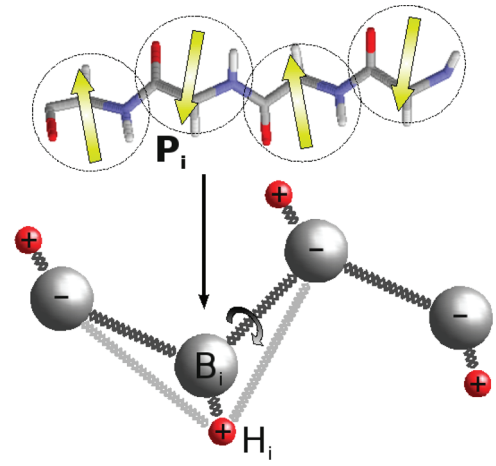


Figure 8. Modeling the protein backbone dipole moment, orientation, and fluctuations using classical Drude-like oscillators. Each particle H_i is bonded to the backbone bead B_i. Assuming that the two particles carry charges of ±0.5*e*, the length of the B_i—H_i bond is equal to $2P_i$, where P_i is the backbone dipole moment. Its orientation relative to the two neighboring backbone beads B_{i-1} and B_{i+1} is determined by the two distances B_{i-1}—H_i and H_i—B_{i+1}.

the dipole of the amino acid (Figure 8). The addition of the particle H_i follows in spirit the classical Drude oscillator model introduced in recent empirical force fields to describe the atomic polarizability.[51,52] In the present model, the auxiliary particles H_i are used to account for both the permanent dipole and the induced polarization of backbone beads.

To determine the backbone dipole moment and orientation relative to the neighboring residues, a statistical analysis was performed on the previous set of experimental structures of all-atom proteins, using the atomic partial charges from the second-generation Amber force field.[53] Assuming arbitrarily that the two grains B_i and H_i carry charges of ±0.5*e*, the analysis yields the probability distribution functions for the B_i—H_i bond length, as well as the B_{i-1}—B_i—H_i and H_i—B_i—B_{i+1} pseudobends (Figure 9). Then, applying the Boltzmann inversion generates the potentials that allow the moment, orientation, and fluctuations of the backbone dipoles to be modeled. As seen in Figure 9, the B_i—H_i bond potential can be satisfactorily fitted with a quadratic function, whereas the two Urey−Bradley-like potentials for the B_{i-1}—B_i—H_i and H_i—B_i—B_{i+1} bends are better fitted with two double-well functions (eq 3), which account for the two preferential α and extended β conformations. It should be noted that the averaged length of the B_i—H_i bonds is equal to 1.4 Å, which reflects a mean dipole moment of 0.7 eÅ for the backbone beads. This value is different from the dipole moment of the peptide bond NH—CO that links two successive Cα atoms (about 2.3 eÅ), because each backbone bead encompasses an amine group NH preceding a CαH and a carbonyl group CO succeeding it. Therefore, in contrast to the planar and quasirigid peptide group in which the bonds NH and CO dipoles are almost colinear and are added, the backbone bead dipole depends on the internal spatial distribution of the NH—CαH—CO atoms, and its moment is less strong and more variable than that of the peptide bond. Nevertheless,

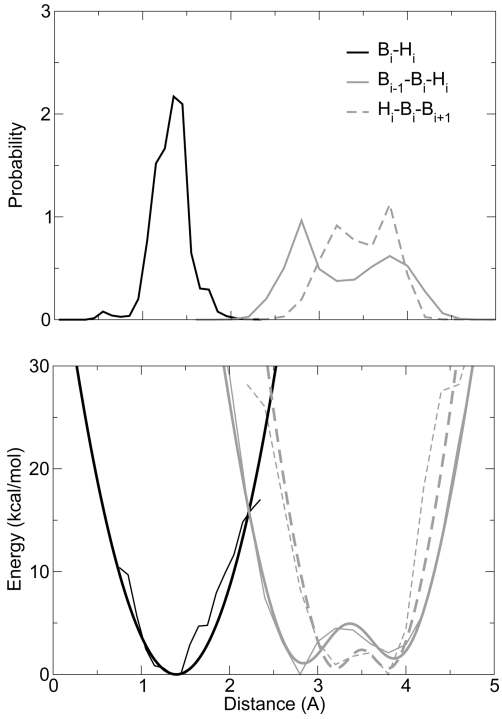


Figure 9. Top: Probability distribution functions for the B_i—H_i bond distance and the B_{i-1}—B_i—H_i and H_i—B_i—B_{i+1} pseudobends. Bottom: Associated potentials (thin lines) and fitting energy functions (thick lines).

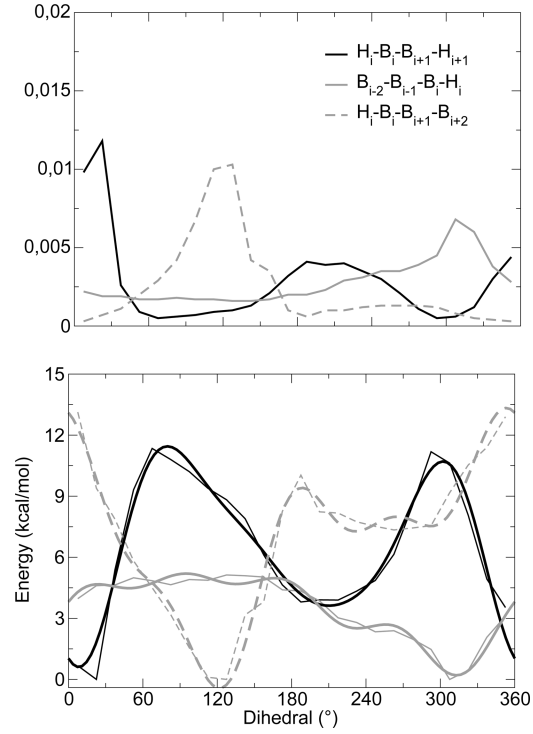


Figure 10. Top: Probability distribution functions for the H_i—B_i—B_{i+1}—H_{i+1}, B_{i-2}—B_{i-1}—B_i—H_i, and H_i—B_i—B_{i+1}—B_{i+2} pseudotorsions. Bottom: Associated potentials (thin lines) and fitting energy functions (thick lines).

the backbone bead dipole mainly arises from the carbonyl group dipole (which is twice as strong as the NH bond), so that, in extended β-strands, where the dipoles of the two NH and CO bonds seem to cancel out, the total backbone bead dipolar moment remains around 0.6 eÅ whereas it has a slightly higher value of around 0.8 eÅ in α-helices.

In most all-atom force fields, the 1−4 nonbonded interactions are treated differently from the others, because they are partially taken into account through dihedral potentials. Similar considerations for the CG model result in the introduction of three additional torsion potentials to account for the dipolar interactions of the B_i bead with its neighboring beads $B_{i\pm1}$ and $B_{i\pm2}$, which interact only through bonded potentials. Figure 10 displays the probability distributions and energy functions for these H_{i-1}—B_{i-1}—B_i—H_i, B_{i-2}—B_{i-1}—B_i—H_i, and H_i—B_i—B_{i+1}—B_{i+2} additional torsion types. Similarly to the B_{i-2}—B_{i-1}—B_i—S_i torsions, the B_{i-2}—B_{i-1}—B_i—H_i energy potential, which has a single minimum, allows for the maintenance of the "chiral" geometry of the substituents of the backbone beads B_i and prevents possible "flips" of the particles H_i. Finally, it can be noticed that this approach locates the dipole of the backbone grains at their geometric center in contrast to the model of Liwo et al.[16] which accounts for the peptide bond dipoles and then locates them in the middle of two successive Cα atoms. Thus, contrary to their dipolar model, the B_i—H_i dipole orientation is not explicitly correlated to the B_{i-2}—B_{i-1}—B_i—B_{i+1} dihedral conformation.[54] Nevertheless, the preferential orientation of the backbone dipoles relative to their neighboring residues is favored by the two torsion potentials H_i—B_i—B_{i+1}—B_{i+2} and H_{i-1}—B_{i-1}—B_i—H_i. All of the parameters for the potentials involving the auxiliary particle H_i are listed in Tables 2−4 of the Supporting Information.

3. Results

The CG bonded potentials described in the preceding section have been combined with the previously developed nonbonded van der Waals potentials (eq 1) in an MD algorithm to test their capability to yield stable trajectories of a few peptides and small proteins. In particular, an important question is the extent to which the flexible reduced protein model preserves the secondary structures and accounts for the loop mobility.

3.1. Simulation Computational Details. The MD simulations of the tested polypeptides were performed with the program Orac,[55] in the canonical *NVT* ensemble using the Nosé−Hoover thermostat[56] and an integration time step of 1 fs. All systems were first progressively heated from 100 to 300 K for 0.6 ns, then equilibrated at this latter temperature for an additional 1.2 ns, and finally simulated for 198.2 ns without any constraint. No cutoff distance was used, so that all nonbonded interactions were taken into account. The present study does not include any explicit hydration model. Nevertheless, to mimic the screening effect of the high-dielectric solvent, all Coulombic interactions were damped using a sigmoidal distance-dependent dielectric function.[57,58] The dielectric screening effect is theoretically dependent on the charge exposure,[59,60] but to keep the solvent model as simple as possible, only two different sigmoidal functions were used: one with a large slope for interactions involving charged side chains, which are generally exposed to the solvent, and one with a smaller slope for interactions between

Table 1. Polypeptides Whose Dynamic Structures Were Simulated with the CG Model

PDB code	no. of residues	nos. of atoms/grains	motif(s)	AA/CG CPU time ratio
1E0Q[68]	17	285/56	β	26.0
2FLY[69]	20	356/70	α	20.5
1VII[64]	36	596/123	α	16.1
1PPT[70]	36	581/119	α/β	19.0
1PGB[71]	56	855/183	α/β	16.7
1BPI[72]	58	892/190	α/β	15.8
1B75[73]	94	1533/319	α/β	17.5
1AJ3[65]	98	1596/334	α	16.7

the backbone beads B_i and H_i, which are overall more buried sites (see Figure 1 in the Supporting Information).

3.2. Equilibrium Structural Properties. Here are reported the results of the MD simulations for eight polypeptides, with sequence lengths ranging from 17 to 98, having various secondary-structure motifs. Table 1 lists the tested peptides and shows that the CG model allows the computational times to be reduced by a factor of about 15−25, relative to the AA description (without explicit solvent). All simulations were started from the protein experimental structure available in the Protein Data Bank,[47] without any addition of counterions. As shown in Figure 2 of the Supporting Information for four of the studied polypeptides, the time evolution of the total energies and their bonded contributions V_{bonded} are stable along the last nanoseconds of the trajectories, suggesting that simulations have reached some equilibrium states. Figure 11 displays the time evolutions of the protein root-mean-square deviations (RMSDs) from their initial conformation, calculated over their backbone beads. Overall, it is observed that most of the tested

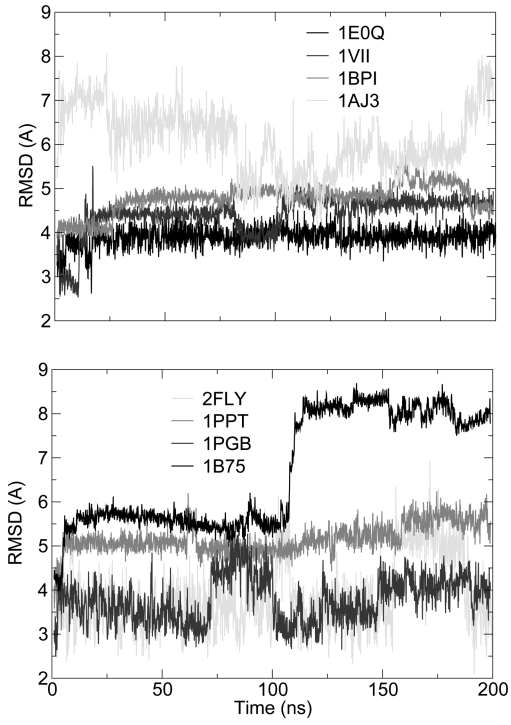


Figure 11. Time evolutions of the RMSD values (calculated over the backbone beads) relative to the initial conformations for the polypeptides (top) 1E0Q, 1VII, 1BPI, and 1AJ3 and (bottom) 2FLY, 1PPT, 1PGB, and 1B75.

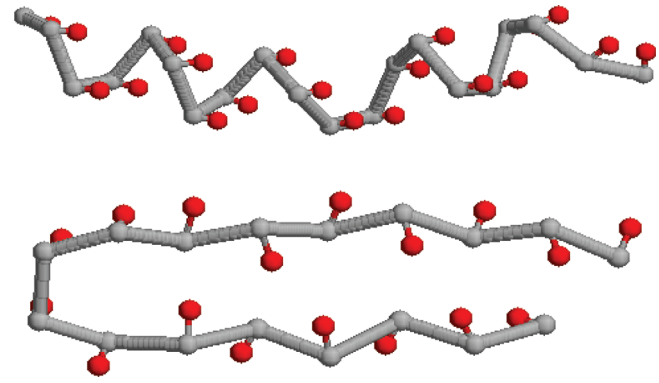


Figure 12. Ball-and-stick representations of the backbone of (top) the α-helix 2FLY and (bottom) the β-hairpin 1E0Q at the end of simulations. The B_i grains are displayed in gray, and the H_i particles are in red.

polypeptides reach stable conformations not too far from the native structures, with RMSDs ranging from 3 to 6 Å after 200 ns of simulation. In contrast, the protein 1B75 structure is quite stable until about 110 ns, after which its RMSD rapidly increases and stabilizes around 8 Å. The protein 1AJ3 seems to have reached some stable conformations between 25 and 75 ns and between 130 and 180 ns, but its RMSD increases at the end of the simulation to a rather high value between 7 and 8 Å. As shown in Figure 3 of the Supporting Information and as discussed below, the final RMSD values for these two proteins clearly indicate large deformations of their tertiary structures. Finally, it can be noticed that the CG bonded potentials do not prevent the proteins, such as 2FLY or 1PGB, to transitionally visit conformations other than the equilibrium one.

Comparisons of these first results with similar CG protein models were difficult, as most of the latter were designed to treat the folding problem and were not used to test the protein stability by MD simulations. Other residue-scale force fields, including the model of Tozzini and McCammon[61] and the MARTINI force field,[24] were developed to study the dynamics properties of large CG proteins over long MD trajectories.[3,62,63] However, as stressed by the authors, their CG force fields were biased to retain either the secondary or the native tertiary structures of proteins, making comparisons with the presented study very delicate. The dynamic conformations generated by their models are certainly closer to the experimental structures than those yielded by ours, as it is generally difficult to have a completely unbiased and accurate force field at the same time. Compared to the recent study of Majek and Elber,[38] which uses an unbiased one-bead backbone model to simulate protein conformational stability, the equilibrium RMSD values displayed in Figure 11 are close to the mean value (5 Å) of the RMSD yielded by their MD simulations. This suggests that the CG force field presented here, which is less elaborate than theirs but similarly models the bonded potentials and hydrogen bonding, allows for the generation of dynamic conformations of simplified proteins as far to the native structures as their simulations.

As illustrated in Figure 12 for the peptides 2FLY and 1E0Q (as well as in Figure 3 of the Supporting Information for the proteins 1PGB and 1AJ3), the polypeptide secondary

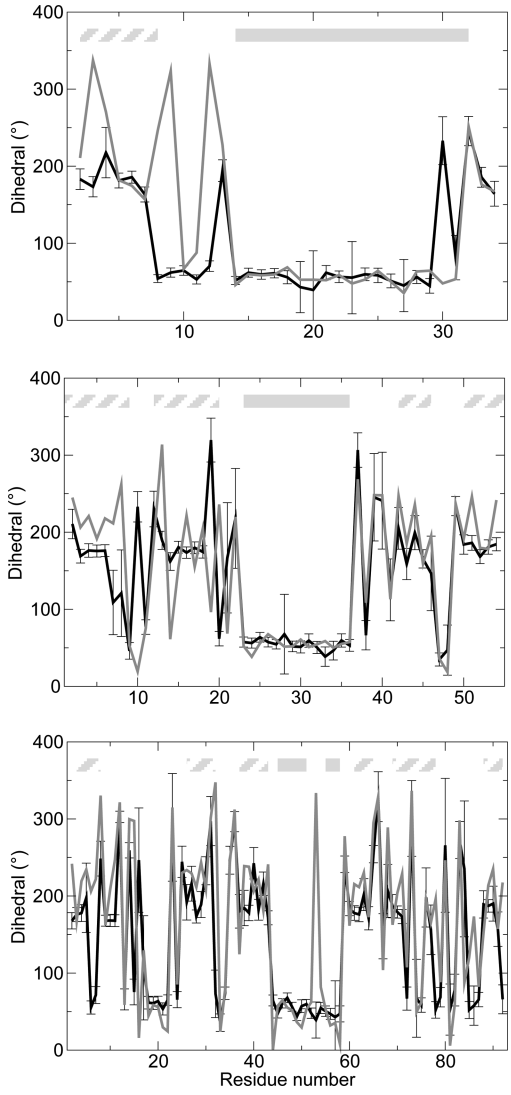


Figure 13. Residue-averaged values of the B_{i-1}—B_i—B_{i+1}—B_{i+2} torsions for the proteins (top) 1PPT, (middle) 1PGB, and (bottom) 1B75. Black and gray lines represent the theoretical predictions and experimental measurements, respectively. Horizontal plain and striped bars indicate the α-helix and β-strand positions, respectively, in the protein sequence.

structures are maintained during the simulations through the electrostatic interactions between backbone beads: In helical chains, the B_i—H_i bonds are mainly oriented along the helix axis, and the grains B_i strongly interact with the H_{i+3} and H_{i+4} particles. In β-sheets, the B_i—H_i dipoles are almost coplanar, perpendicular to the strand direction and alternatively point toward and from the neighboring strand. Without the simplified model of hydrogen bonds, the α-helix 2FLY bends rapidly, and a kink occurs in its middle, close to the Gly residue, and in the β-hairpin simulations, the two strands locally move apart, become highly curved, and dramatically lose their parallel orientation.

The values of the backbone torsions B_{i-1}—B_i—B_{i+1}—B_{i+2} provide a quantitative indicator of the stabilities of the secondary structures during the MD simulations.[18] Figure 13 displays the averaged values of these dihedral angles along the sequence of polypeptides 1PPT, 1PGB, and 1B75, which have various lengths and a common α/β motif, in order to

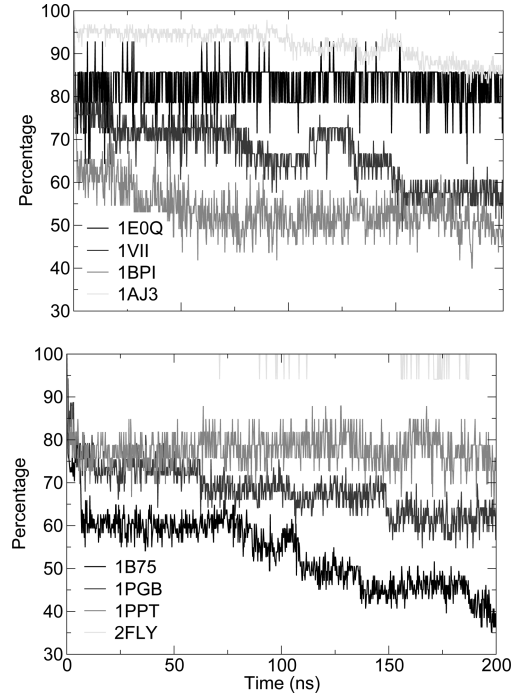


Figure 14. Time evolution of the percentage of native backbone torsions for the polypeptides (top) 1E0Q, 1VII, 1BPI, and 1AJ3 and (bottom) 2FLY, 1PPT, 1PGB, and 1B75.

display the different angle values that correspond to the backbone torsions. When compared with the values measured in the PDB structures, it can be observed that most of the backbone torsions keep a conformation close to the experimental one, especially in the helical and β-strand structures. Several of them undergo transitions, and such conformational changes mainly occur near or in the loop regions. In the case of 1PPT, an α → β transition of the residue 30 backbone torsions can be observed that slightly kinks the helix end, as well as a structural change into an α-conformation of the loop region 8−12, not observed by experiments. In the protein 1B75 simulation, several dihedral angles undergo β → α transitions, particularly in the first and fifth β-strands, revealing large conformational rearrangements consistent with the relatively high RMSD time evolution. It should be emphasized here that, whereas the residues in α-helices have dihedral values around 70°, which is the angle value of the first minimum in the B_{i-1}—B_i—B_{i+1}—B_{i+2} potential (Figure 5), the residues in β-strands have average torsions around 190°, which does not correspond to the second minimum of the backbone potential. These β conformations are, in fact, partially stabilized by the H_i—B_i—B_{i+1}—H_{i+1} torsions (Figure 10), as well as by the S_i—B_i—B_{i+1}—B_{i+2} potentials (Figure 7), which are sensitive to the nature of the amino acids. After many tests, no stable structure of proteins could be obtained without the H_i—B_i—B_{i+1}—H_{i+1} potential. Overall, the CG bonded potentials allow the backbone dihedral conformation of the proteins to be well conserved: As shown in Figure 14, which displays a plot of the time evolutions of the percentage of native backbone torsions (plus or minus 30°), most of the MD trajectories preserve more than 60% of the protein backbone dihedral angles in the experimental conformation. The two simulations of proteins 1BPI and 1B75

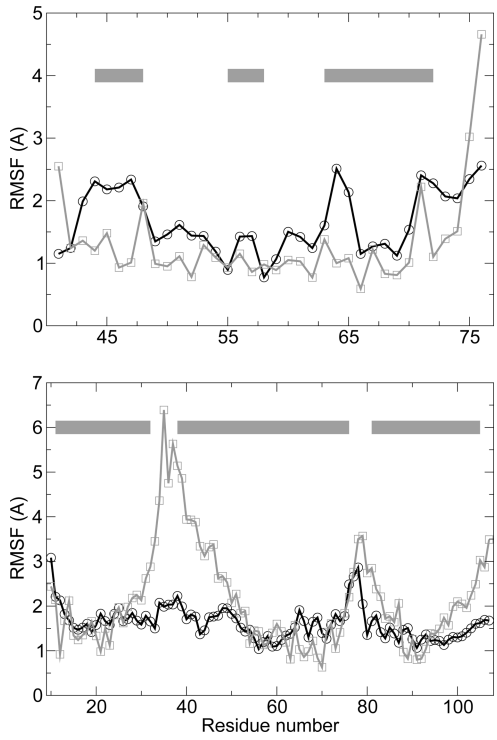


Figure 15. Comparison of the residue RMSFs relative to averaged conformations, computed from simulations (black lines) or provided by NMR experiments (gray lines), for the proteins (top) 1VII and (bottom) 1AJ3. Horizontal dark gray and light gray lines represent the α-helix positions.

are those that least conserve the backbone structures, with a percentage of torsion values close to experiments of around 50%.

If the secondary structures are relatively well conserved along the MD simulations, some protein tertiary structures are slightly deformed relative to the experimental structures. For example, the two last α-helices of the protein 1AJ3 are more curved and twisted at the end of the simulation than in the initial conformation (see Figure 3 of the Supporting Information). During the simulation of 1VII, the orientation of its first α-helix relative to the other two can change significantly. At the end of the 1PGB simulation, the α-helix is oriented almost parallel to the β-strand direction, whereas it diagonally crosses the sheets in the experimental structure. In the protein 1B75, the initial conformation in the β-barrel become flatter instead of remaining roughly cylindrical. All of these discrepancies could be due to the absence of counterions and of an electrostatic description of the neutral polar side chains. They probably also arise from the crude model of solvation, which cannot account for hydrophobic interactions within the protein cores or for the molecular nature and exclusion-volume effect of the solvent. Further corrections of these deficiencies in the CG protein model are at present under development.

3.3. Structural Fluctuations and Thermal Instability. One question addressed here is whether the simplified backbone model can reasonably simulate the conformational fluctuations of proteins, particularly their loops, which are expected to undergo the largest-amplitude movements. In Figure 15 are plotted the residue root-mean-square fluctuations (RMSFs) relative to the averaged structures for the 1VII and 1AJ3 proteins. These RMSFs are directly compared to those provided by NMR experiments that have solved the three-dimensional structures.[64,65] The RMSFs calculated for 1VII appear larger than the NMR values, especially for residues in the first α-helix and at the beginning of the third α-helix. This confirms that the CG protein tertiary structure is less compact than the experimental structure and thus does not restrain the movements of its secondary structures much. In contrast, the 1AJ3 loops undergo fewer large-amplitude motions than in NMR experiments, especially the loop connecting the first two α-helices. When examining the CG protein structure (Figure 3 of the Supporting Information), it can be noticed that the loops come close to and interact with the C-terminus and N-terminus instead of being fully solvated, which could explain their rather restrained motions. Nevertheless, it is overall observed that the residues in the secondary structures have the lowest RMSFs, whereas the loop regions have the largest fluctuations, no more nor less accurately than calculations provided by less detailed elastic network models.

A finer indicator of the protein backbone conformational fluctuations is provided by the measurement of the so-called S^2 order parameter with NMR spectroscopy.[66] This parameter reflects the angular mobility of the backbone N—H bonds and can be compared to the angular mobility of the B—H dipoles in the CG protein model. In simulations, it can be calculated as

$$S^2 = \frac{3}{2}\sum_{\alpha=1}^{3}\sum_{\beta=1}^{3}\langle \mu_\alpha \mu_\beta \rangle^2 - \frac{1}{2} \qquad (5)$$

where μ_α and μ_β denote the three components, x, y, and z, of the normalized vector along the B—H bonds and where the brackets symbolize a time average over simulations.[67] An S^2 order parameter close to 1 indicates that the N—H or B—H vectors are quite constrained in space and that the H particle is probably involved in a hydrogen bond. If it is close to 0, these vectors are, on the contrary, free to move and rotate. The S^2 parameters calculated from the simulations are displayed in Figure 16 as a function of the residue numbers of the two proteins 1BPI and 1PGB. Overall, it is observed that the B—H bonds have small angular fluctuations for residues in α-helices and β-strands, whereas they are significantly more mobile for residues in loops. This corroborates the role of the electrostatic interactions between B—H polar groups in stabilizing the secondary structures, similarly to the hydrogen bonds between the N—H and C═O atoms. However, discrepancies between simulations and experiments can be noticed: In the 1BPI simulation, the two β-strands are significantly more mobile, whereas the loop between them has more restrained movements than observed by NMR spectroscopy, indicating a slight deformation of these secondary structures. The S^2 parameter profile of the protein 1PGB is in overall good agreement with experimental observations, except for the third β-strand, which appears to be more mobile, as well as the second loop before the α-helix, which undergoes larger-amplitude motion. Further investigations, especially into the stabilizing role of solvent, are being conducted to explain these discrepancies and

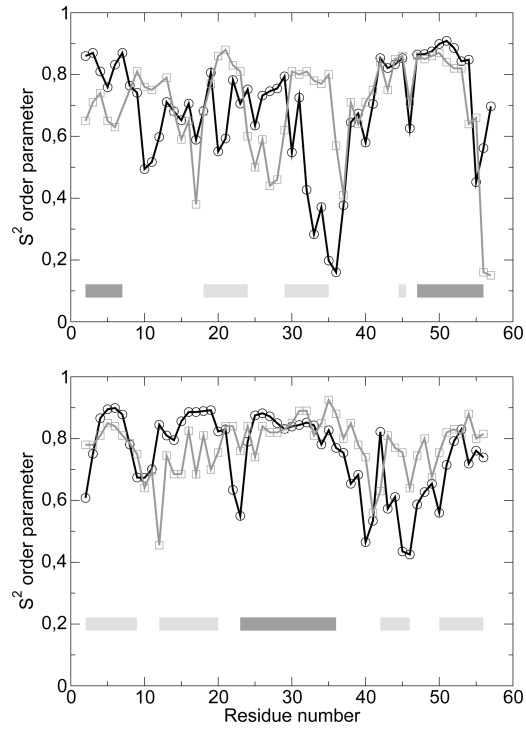


Figure 16. Comparison of the B—H order parameters (black line) with the N—H order parameters measured by NMR spectroscopy (gray line) for the proteins (top) 1BPI and (bottom) 1PGB. Horizontal dark gray and light gray lines represent α-helix and β-strand positions, respectively.

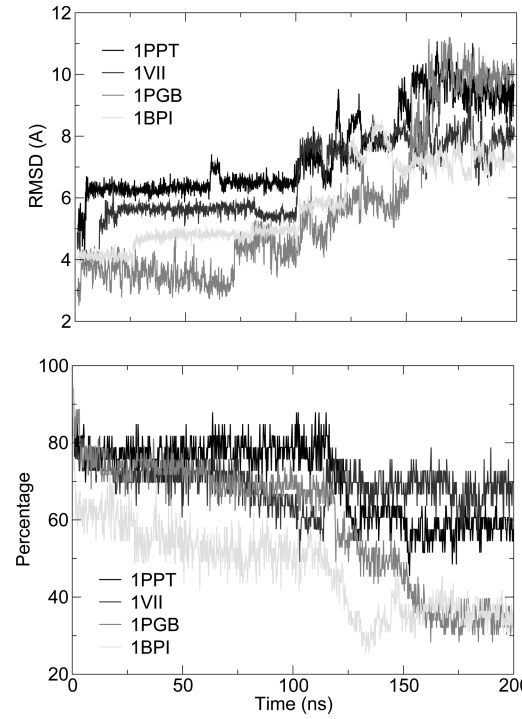


Figure 17. Time evolutions of the (top) RMSD relative to the initial conformation and (bottom) percentage of native backbone torsions for the polypeptides 1PPT, 1VII, 1PGB, and 1BPI. At 100 ns, the temperature was increased from 300 to 400 K, and at 150 ns, it was increased again to 500 K.

improve the capability of the CG model to yield stable and meaningful dynamic conformations of proteins.

As previously mentioned, the CG bonded potentials presented here possibly enable conformation changes of the simulated proteins, contrary to elastic network models, which can study only harmonic deformations around a single structure. To clarify the capability of CG proteins to explore non-native conformations, simulations at high temperatures were performed for the four medium-size proteins 1PPT, 1VII, 1PGB, and 1BPI. In practice, their previous MD trajectories at 300 K were continued at 400 K from 100 to 150 ns then at 500 K from 150 to 200 ns. Figure 17 shows the time evolutions of the RMSDs from the experimental structures and the percentages of native backbone torsions (±30°) for the four simulations. It can be observed that the RMSDs of the four proteins increase during the 50-ns trajectories at 400 K and, for the proteins 1PPT and 1PGB, continue to rise during the last 50 ns at 500 K to values ranging from 7 to 10 Å. This clearly indicates a denaturation of their native conformation upon heating. Nevertheless, regarding the percentages of native backbone torsions, the simulations of the two proteins 1PPT and 1VII maintain a rather high ratio of backbone torsions in native conformation (between 60% and 70%), suggesting that their tertiary structure is lost before their secondary structure. In contrast, the percentages of native backbone torsions of the two protein 1PGB and 1BPI rapidly decrease to low values between 30% and 40%. For these denatured proteins, the secondary structures are lost concomitantly to their tertiary conformation.

4. Conclusions

This work introduces a set of bonded potentials for modeling the backbone flexibility of proteins described with residue-scale coarse grains. The main feature of these potentials is that the local secondary-structure propensity of the amino acids seems to be encoded in the pseudotorsions S_i—B_i—B_{i+1}—B_{i+2}. Combined with the nonbonded van der Waals potential reported recently[29] and a simplified dipolar model of hydrogen bonds, the CG bonded potentials overall generate stable protein structures in the neighborhood of the experimental conformations. Despite some backbone torsion conformational changes, the protein secondary structures are quite well conserved along the 200-ns MD trajectories, but some tertiary structures deviate from the initial ones, especially the β-barrel protein 1B75. The amino acids in the loops connecting helices and strands are found to have the largest internal mobility. At present, further improvements of the model, particularly a better description of the polar residues and their interactions with high-dielectric solvents, are in development. It is believed that this CG protein model will provide, at a low computational cost, a reasonable dynamic picture of protein equilibrium structures and useful insights into the functional role of large protein conformational changes.

Acknowledgment. The author acknowledges Grant ANR-08-JCJC-0081-01 from the Agence Nationale de la Recherche.

Supporting Information Available: All of the parameters of the coarse-grained protein bonded potentials de-

scribed herein. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Henzler-Wildman, K.; Kern, D. *Nature* **2007**, *450*, 964.

(2) Ishima, R.; Freedberg, D.; Wang, Y.; Louis, J.; Torchia, D. *Structure* **1999**, *7*, 1047.

(3) Trylska, J.; Tozzini, V.; Chang, C.; McCammon, J. *Biophys. J.* **2007**, *92*, 4179.

(4) Gunasekaran, K.; Ma, B.; Nussinov, R. *Proteins* **2004**, *57*, 433.

(5) Popovych, N.; Sun, S.; Ebright, R.; Kalodimos, C. *Nat. Struct. Mol. Biol.* **2006**, *13*, 831.

(6) Ishima, R.; Torchia, D. *Nat. Struct. Biol.* **2000**, *7*, 740.

(7) Karplus, M.; McCammon, J. *Nat. Struct. Biol.* **2002**, *9*, 646.

(8) Muller-Plathe, F. *Chem. Phys. Chem.* **2002**, *3*, 755.

(9) Zhou, J.; Thorpe, I.; Izvekov, S.; Voth, G. *Biophys. J.* **2007**, *92*, 4289.

(10) Klein, M.; Shinoda, W. *Science* **2008**, *321*, 798.

(11) Durrieu, M.; Bond, P.; Sansom, M.; Lavery, R.; Baaden, M. *Chem. Phys. Chem.* **2009**, *10*, 1548.

(12) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59.

(13) Kolinski, M.; Skolnick, J. *Polymer* **2004**, *45*, 511.

(14) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

(15) Bonvin, A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194.

(16) Liwo, A.; Pincus, M.; Wawak, R.; Rackovsky, S.; Scheraga, H. *Protein Sci.* **1993**, *2*, 1715.

(17) Wallqvist, A.; Ullner, M. *Proteins* **1994**, *18*, 267.

(18) DeWitte, R.; Shakhnovich, E. *Protein Sci.* **1994**, *3*, 1570.

(19) Reva, B.; Finkelstein, A.; Sanner, M.; Olson, A. *Protein Eng.* **1997**, *10*, 865.

(20) Bahar, I.; Kaplan, M.; Jernigan, R. *Proteins* **1997**, *29*, 292.

(21) Haliloglu, T.; Bahar, I. *Proteins* **1998**, *31*, 271.

(22) Derreumaux, P. *J. Chem. Phys.* **1999**, *111*, 2301.

(23) Van Giessen, A.; Straub, J. *J. Chem. Theory Comput.* **2006**, *2*, 674.

(24) Monticelli, L.; Kandasamy, S.; Periole, X.; Larson, R.; Tieleman, D.; Marrink, S. *J. Chem. Theory Comput.* **2008**, *4*, 819.

(25) Gabdoulline, R.; Wade, R. *J. Phys. Chem.* **1996**, *100*, 3868.

(26) Reith, D.; Putz, M.; Muller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624.

(27) Izvekov, S.; Voth, G. *J. Chem. Phys.* **2005**, *123*, 134105.

(28) Prampolini, G. *J. Chem. Theory Comput.* **2006**, *2*, 556.

(29) Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Phys. Chem. B* **2007**, *111*, 9390.

(30) Tirion, M. *Phys. Rev. Lett.* **1996**, *77*, 1905.

(31) Bahar, I.; Atilgan, A.; Erman, B. *Fold. Des.* **1997**, *2*, 173.

(32) Hinsen, K. *Proteins* **1998**, *33*, 417.

(33) Tama, F.; Sanejouand, Y. *Protein Eng.* **2001**, *14*, 1.

(34) Micheletti, C.; Carloni, P.; Maritan, A. *Proteins* **2004**, *55*, 635.

(35) Klimov, D.; Betancourt, M.; Thirumalai, D. *Fold. Des.* **1998**, *3*, 481.

(36) Voegler Smith, A.; Hall, C. *Proteins* **2001**, *44*, 344.

(37) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. *J. Chem. Phys.* **1999**, *110*, 11616.

(38) Majek, P.; Elber, R. *Proteins* **2009**, *76*, 822.

(39) Yap, E.; Fawzi, N.; Head-Gordon, T. *Proteins* **2008**, *70*, 626.

(40) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. *J. Phys. Chem. B* **2005**, *109*, 13798.

(41) Liwo, A.; Khalili, M.; Scheraga, H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362.

(42) Miyazawa, S.; Jernigan, R. *Macromolecules* **1985**, *18*, 534.

(43) Sippl, M. *J. Mol. Biol.* **1990**, *213*, 859.

(44) Bryant, S.; Lawrence, C. *Proteins* **1993**, *16*, 92.

(45) Skolnick, J.; Jaroszewski, L.; Kolinski, A.; Godzik, A. *Protein Sci.* **1997**, *6*, 676.

(46) Betancourt, M.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 361.

(47) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Res.* **2000**, *28*, 235.

(48) Kozlowska, U.; Liwo, A.; Scheraga, H. *J. Phys. Condens. Matter* **1997**, *19*, 285203.

(49) Liwo, A.; Pincus, M.; Wawak, R.; Rackovsky, S.; Oldziej, S.; Scheraga, H. *J. Comput. Chem.* **1997**, *18*, 874.

(50) Chou, P.; Fasman, G. *Biochemistry* **1974**, *13*, 211.

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

(52) Vorobyov, I.; Anisimov, V.; MacKerell, A., Jr. *J. Phys. Chem. B* **2005**, *109*, 18988.

(53) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K., Jr.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(54) Liwo, A.; Oldziej, S.; Czaplewski, C.; Kozlowska, U.; Scheraga, H. *J. Phys. Chem. B* **2004**, *108*, 9421.

(55) Procacci, P.; Darden, T.; Paci, E.; Marchi, M. *J. Comput. Chem.* **1997**, *18*, 1848.

(56) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511.

(57) Mehler, E.; Eichele, G. *Biochemistry* **1984**, *23*, 3887.

(58) Hingerty, B.; Ferrell, T.; Turner, J. *Biopolymers* **1985**, *24*, 427.

(59) Sandberg, L.; Edholm, O. *Proteins* **1999**, *36*, 474.

(60) Mallik, B.; Masunov, A.; Lazaridis, T. *J. Comput. Chem.* **2002**, *23*, 1090.

(61) Tozzini, V.; McCammon, J. *Chem. Phys. Lett.* **2005**, *413*, 123.

(62) Shih, A.; Arkhipov, A.; Freddolino, P.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674.

(63) Treptow, W.; Marrink, S.; Tarek, M. *J. Phys. Chem. B* **2008**, *112*, 3277.

(64) McKnight, C.; Matsudaira, P.; Kim, P. *Nat. Struct. Biol.* **1997**, *4*, 180.

(65) Pascual, J.; Pfuhl, M.; Walther, D.; Saraste, M.; Nilges, M. *J. Mol. Biol.* **1997**, *273*, 740.

(66) Barchi, J., Jr.; Grasberger, B.; Gronenborn, A.; Clore, G. *Protein Sci.* **1994**, *3*, 15.

(67) Smith, P.; van Schaik, R.; Szyperski, T.; Wuthrich, K.; van Gunsteren, W. *J. Mol. Biol.* **1995**, *246*, 356.

(68) Zerella, R.; Chen, P.; Evans, P.; Raine, A.; Williams, D. *Protein Sci.* **2000**, *9*, 2142.

(69) Lucyk, S.; Taha, H.; Yamamoto, H.; Miskolzie, M.; Kotovych, G. *Biopolymers* **2006**, *81*, 295.

(70) Blundell, T.; Pitts, J.; Tickle, I.; Wood, S.; Wu, C. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 4175.

(71) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. *Biochemistry* **1994**, *33*, 4721.

(72) Parkin, S.; Rupp, B.; Hope, H. *Acta Crystallogr. D* **1996**, *52*, 18.

(73) Stoldt, M.; Wohnert, J.; Gorlach, M.; Brown, L. *Embo J.* **1998**, *17*, 6377.

CT900408S

JCTC Journal of Chemical Theory and Computation

Simulating Monovalent and Divalent Ions in Aqueous Solution Using a Drude Polarizable Force Field

Haibo Yu,[†,‡] Troy W. Whitfield,[‡,§,#] Edward Harder,[†] Guillaume Lamoureux,[||] Igor Vorobyov,[⊥,▽] Victor M. Anisimov,[⊥] Alexander D. MacKerell, Jr.,[⊥] and Benoît Roux*[,†,§]

Department of Biochemistry and Molecular Biology, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60649, Department of Chemistry and Biochemistry, Concordia University, Montréal, Québec, H4B 1R6, Canada, and Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201

Received October 29, 2009

Abstract: An accurate representation of ion solvation in aqueous solution is critical for meaningful computer simulations of a broad range of physical and biological processes. Polarizable models based on classical Drude oscillators are introduced and parametrized for a large set of monatomic ions including cations of the alkali metals (Li^+, Na^+, K^+, Rb^+, and Cs^+) and alkaline earth elements (Mg^{2+}, Ca^{2+}, Sr^{2+}, and Ba^{2+}) along with Zn^{2+} and halide anions (F^-, Cl^-, Br^-, and I^-). The models are parametrized, in conjunction with the polarizable SWM4-NDP water model [Lamoureux et al. *Chem. Phys. Lett.* **2006**, *418*, 245], to be consistent with a wide assortment of experimentally measured aqueous bulk thermodynamic properties and the energetics of small ion−water clusters. Structural and dynamic properties of the resulting ion models in aqueous solutions at infinite dilution are presented.

1. Introduction

Ions are fundamental to the structure and function of biological systems, where their local environment can be as diverse as are the roles that they play. Ions are involved in the folding of proteins and nucleic acids, enzyme catalysis, and numerous cellular signaling processes. Monovalent ions such as Na^+, K^+, and Cl^- play an important role in the homeostasis and electric activity of living cells and in modulating biomolecular stability through both specific and nonspecific interactions.[1,2] Divalent cations, such as Zn^{2+}, Mg^{2+}, and Ca^{2+}, are often associated with catalytic or regulatory activities of proteins and enzymes,[3,4] including the activation by Ca^{2+} of K^+ channels,[5] the stabilization by Zn^{2+} of zinc-finger proteins,[6] and the condensation and folding of RNA and DNA by Mg^{2+}.[7−12]

In order to have meaningful computational models to probe and explore the diverse and important roles of ions in biological phenomena, accurate and physically realistic descriptions of their microscopic interactions is crucial. This is not trivial because a balanced description of the interactions between ion and ion, ion and water, and ion and biomolecules is required. A number of nonpolarizable models for ions have been developed.[13−18] The influence of induced electronic polarization, in particular, has been shown to be critical to the study of ion channels,[19−21] and it is expected to be generally important for the study of aqueous ionic systems. Striking a balance between accuracy and computational expense is an important consideration in designing useful models. Because proper statistical averaging in biological

* Corresponding author e-mail: roux@uchicago.edu.
† The University of Chicago.
§ Argonne National Laboratory.
|| Condordia University.
⊥ University of Maryland.
‡ These authors contributed equally to this paper.
# Current address: Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcestor, Massachusetts 01605.
▽ Current address: Department of Chemistry, University of California, Davis, California 95616.

10.1021/ct900576a
Published on Web 01/29/2010

systems requires long simulations to sample over many configurations,[22] achieving a sufficient sampling of the relevant configurations can become computationally prohibitive if the microscopic interactions are generated via sophisticated ab initio quantum mechanical (QM) approaches. In that regard, treatments of many-body polarization effects based on simple potential functions have important advantages over QM methods. An alternative approach is to adopt a polarizable force field.[23−26]

Generally speaking, there are three different methods to account for explicit electronic polarization into classical force fields, i.e., induced dipole model, fluctuation charge model, and classical Drude oscillator model. Correspondingly, a number of polarizable models for ions have been developed based on these three different approaches to study a variety of phenomena involving ions.[27−30] Our own efforts have been focused on developing a complete all-atomic polarizable force field for proteins, nucleic acids, and membranes based on the concept of the classical Drude oscillator.[31] This model approximates the quantum mechanical electronic responses by using auxiliary massless charged particles that are harmonically bound to the polarizable nuclei.[24,26,29,32−35] This approach is also referred to as the "shell model"[36−38] or "charge-on-spring".[24,39] The polarizable Drude force field has been shown in previous studies to be accurate for the simulation of liquid water,[33−35] aqueous ionic systems,[29] condensed phases of small organic molecules,[40−44] and lipid membranes.[45] In these efforts, all new chemical entities must be compatible with the polarizable SWM4-NDP water model,[35] which serves as the central reference for the polarizable Drude force field.

In the present paper, models for an extended set of ions are constructed and parametrized to be compatible with the SWM4-NDP water model.[35] The set comprises the most abundant and biologically relevant monatomic ions, including cations of the alkali metals (Li^+, Na^+, K^+, Rb^+, and Cs^+) and alkaline earth elements (Mg^{2+}, Ca^{2+}, Sr^{2+}, and Ba^{2+}) along with Zn^{2+} and halide anions (F^-, Cl^-, Br^-, and I^-). The present models of alkali and halide ions are similar to those developed previously,[29] which were parametrized in conjunction with the SWM4-DP water model.[34]

Because the SWM4-DP and SWM4-NDP models differ in the sign of the charge on the auxiliary Drude particle attached to the oxygen atom, a reparametrization is necessary to have models compatible with the SWM4-NDP water model.[35] In addition, models for divalent cations are introduced and parametrized. The latter pose new challenges for a classical polarizable force field, particularly regarding the treatment of overpolarization and Coulombic singularities. Approaches to overcome these problems are presented.

The paper is organized as follows. Section 2 recapitulates the relevant details of the Drude model for ion solvation and articulates the questions that arise when extending it to treat highly polarizable anions and divalent ions, along with steps that have been taken to solve the problem. Also included in section 2 are the parametrization strategy and target data. In section 3, results of the fitting are presented for our set of ions, along with an analysis of the structural properties and polarization effects that are predicted by the model.

2. Theory and Methods

2.1. Classical Drude Polarizable Model. The model ions are consistent with the SWM4-NDP polarizable water model with a negatively charged Drude oscillator bound to its oxygen site.[35] The SWM4-NDP potential reproduces most properties of bulk water under ambient conditions (density, vaporization enthalpy, radial distribution function, dielectric constant, self-diffusion constant, shear viscosity, and free energy of hydration). In particular, the SWM4-NDP model yields a correct static dielectric constant, which makes it appropriate to study systems dominated by water-mediated electrostatic interactions. The SWM4-NDP water model comprises five interaction sites: (1) the oxygen atom "O" carrying a charge of $-q_D^{wat}$, (2) the Drude particle "D" harmonically attached to the oxygen atom carrying a (negative) charge of q_D^{wat}, (3) the massless auxiliary site "M" carrying a charge q_M^{wat}, and (4 and 5) the hydrogen atoms"H_1" and "H_2", each carrying a charge of q_H^{wat}. The interactions within and between water molecules are calculated according to the formulation of the polarizable Drude model described in refs 33 and 34. The ion models comprise two sites: (1) the ion core atom "A" carrying a charge of $q^{ion} - q_D^{ion}$ and (2) a Drude particle "D" attached to the ion atom carrying a (negative) charge of q_D^{ion}. To be consistent with the SWM4-NDP model, polarization of an ion is represented by a negatively charged particle, representing electron density, bound to its nucleus (core). All atomic dispersion and electronic overlap effects are represented in a pairwise additive way using the Lennard−Jones (LJ) potential. The core repulsion and van der Waals dispersive interactions are modeled by a LJ interaction between the water oxygen and the ion core atom via the Lorentz−Berthelot combination rule.[46] The potential energy of one ion and one water molecule can be written as

$$U_{tot}^{iw} = U_{LJ}^{iw} + U_{Drude}^{iw} + U_{elec}^{iw} \quad (1)$$

where U_{LJ}^{iw} is the LJ interaction between the ion and the water oxygen atom, U_{Drude}^{iw} represents the harmonic restoring springs associated with the Drude oscillator of the water molecule and the ion, and U_{elec}^{iw} includes all the Coulombic electrostatic interactions between the fixed and mobile charges carried by the ion (two sites) and the water molecule (five sites). The spring constant K_D is set to 1000 kcal/mol/Å^2 for all Drude oscillators in the system. This value dictates the magnitude of the charge that the Drude particle must carry to produce the correct polarizability α, i.e., $q_D = -(\alpha K_D)^{1/2}$.[35] For example, the charge on the Drude particle of I^-, the most polarizable ion in the current study, is $-4.733e$ with this restoring spring constant.

Simulations of the models are performed by considering the dynamics of an extended Lagrangian in which a small mass m_D and kinetic energy are attributed to the Drude particles. The amplitude of the Drude oscillators is controlled with a low-temperature thermostat acting in the local center-of-mass reference frame of each atom−Drude pair.[33] The mass of the Drude particles is set to 0.4 amu, which is subtracted from the mass of the physical polarizable nucleus such that the total mass of the pair remains constant. To

ensure that the time course of the induced dipoles stays close to the self-consistent field (SCF) solution, a Nosé−Hoover thermostat at a temperature $T_* = 1$ K is applied to the relative motion of each atom−Drude pair (in their local center-of-mass reference frame). It was shown that the trajectories generated according to this procedure are very close to those generated by the SCF regime of induced polarization.[33,34] To control the global thermalization of the system, a second thermostat is applied to the center of mass of the atom−Drude pairs as well as the hydrogen atoms.

2.2. Overpolarization of Drude Oscillators. The simple sum over Coulomb interactions of U_{elec} in eq 1 does not exclude singular r^{-1} attractive interactions between the Drude particles and other interaction sites carrying a net charge. Such singularities are generally not problematic in fixed-charge force fields, where the charges are buried within r^{-12} LJ core repulsive interactions. In the polarizable model based upon the Drude oscillator, however, the charge on the Drude particles is not as effectively shielded from other charges by such nonelectrostatic core repulsive interactions. To clarify the situation leading to singularities, consider the interaction between a Drude oscillator bound to a heavy atom fixed at the origin and a point charge q_i placed at some distance X. Along the one-dimensional axis, the potential energy is

$$U(x) = \frac{1}{2}K_D x^2 + \frac{q_D q_i}{|X|} - \frac{q_D q_i}{|X - x|} \quad (2)$$

where x is the displacement of the oscillator, q_D is the magnitude of the charge on each end of the Drude oscillator, and K_D is the harmonic restoring force constant. The self-consistent field condition is

$$\frac{dU(x)}{dx} = K_D x - \frac{q_D q_i (X - x)}{|X - x|^3} = 0 \quad (3)$$

The point of inflection for this solution becomes unstable when

$$X = \left(\frac{27}{4} q_i \sqrt{\frac{\alpha}{K_D}}\right)^{1/3} \quad (4)$$

at which point $x = X/3$. For a fixed value of K_D, instabilities can occur when the polarizability α is large, as in the case of some anions, or when the charge q_i is large, as in the case of the small monatomic divalent cations.

To illustrate the instabilities encountered with a divalent cation ($q_i = 2.0e$), we substitute the charge and force constant parameters from the SWM4-NDP water model. The results are illustrated in Figure 1. Substituting parameters for the SWM4-NDP water model into eq 4, one finds that $X = 1.974$ Å. For example, when the divalent cation is placed at a distance of $2X$ from the origin, the Drude particle can reside in a stable minimum located at $x \approx 0.14$ Å (Figure 1). However, as shown in Figure 1, placing the divalent cation $X < 1.974$ Å will cause the SWM4-NDP oscillator to fall into the singular Coulomb well at X. In contrast, the system is stable with a monovalent cation ($q_i = 1.0e$): according to eq 4, the instability appears only at $X = 1.567$ Å.[29] This distance is considerably shorter than that observed for ion−water close contacts in simulation of ions in solution.

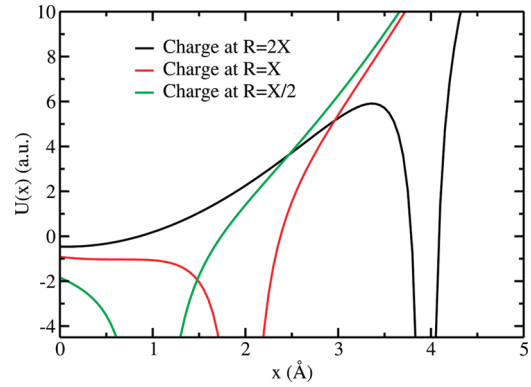


Figure 1. SWM4-NDP water molecule with a point charge of $+2e$ placed $2X$, X, and $X/2$ defined by eq 4 away from the oxygen atom.

Monovalent cations, unlike those for the divalent cations with the oxygen Drude sites on SWM4-NDP water molecules, therefore need no special treatment in order to avoid the "polarization catastrophe".[29]

A number of empirical remedies to the overpolarization problems are possible. The simplest treatment is to introduce an additional anharmonic restoring force to prevent excessively large excursions of the Drude particle away from the atom.[39] The value of such anharmonic term can be adjusted to reduce the polarizability of atoms at high field. This approach was used to prevent any instabilities with the highly polarizable anions such as Br^- and I^-. The anharmonic restoring force was chosen to be active only beyond a certain stretching distance ΔR_{cut}

$$U_{hyp} = K_{hyp} \cdot (\Delta R - \Delta R_{cut})^4 \quad (5)$$

thus preserving the linear polarization response with small displacements.

The cutoff was chosen to be based on the maximum displacement of the Drude particles with respect to their nuclei in SWM4-NDP water models to ensure that its properties will not be affected. The force constant was chosen to be 40 000 kcal/mol/Å^4 to reproduce the maximum induced dipole moment of halide anions estimated by MP2 calculations (see Supporting Information Figure S1). If the stiff anharmonic restoring force causes numerical instability with the finite time-step integrator, the problem could be treated using a multiple time-step algorithm, where the rapidly varying restoring forces are integrated with a smaller time step than the remaining slowly varying forces.[47]

A second possibility is to associate a small repulsive core to the Drude particle implemented with the NBFIX option in the program CHARMM.[48] Finally, another possibility is to introduce electrostatic screening functions that alter the charge−charge $1/r$ Coulombic interactions at short distances.[30,49−52] The latter approach was used to construct stable and accurate models of the divalent cations. A screening function was introduced to modulate the electrostatic interactions between the divalent ion and the induced dipole component of the oxygen of the water molecules

$$\frac{q_i q_j}{r_{ij}} \rightarrow \left(\frac{q_i q_j}{r_{ij}}\right) S_{\mathrm{iw}}(r_{ij}) \tag{6}$$

where r_{ij} are the distances between the pairs of charges taken from $q_i = \{q^{\mathrm{ion}} - q_{\mathrm{D}}^{\mathrm{ion}}, q_{\mathrm{D}}^{\mathrm{ion}}\}$ and $q_j = \{-q_{\mathrm{D}}^{\mathrm{wat}}, q_{\mathrm{D}}^{\mathrm{wat}}\}$. The screening function $S_{\mathrm{iw}}(r)$ is given by

$$S_{\mathrm{iw}}(r) = 1 - \left(1 + \frac{r}{2a_{\mathrm{iw}}}\right) e^{-r/a_{\mathrm{iw}}} \tag{7}$$

where $a_{\mathrm{iw}} = (\alpha_{\mathrm{i}}\alpha_{\mathrm{w}})^{(1/6)}/t_{\mathrm{iw}}$ and α_{i} and α_{w} are the polarizabilities of the ion and oxygen of the water molecule, respectively. This functional form of electrostatic screening was originally introduced for point dipoles by Thole.[40] The dimensionless Thole parameter t_{iw} modulates the strength of the screening for the *ij* pair. In the polarizable Drude model, this form has previously been utilized to modulate the intramolecular nearest-bonded-neighbor 1−2 interactions and next-nearest-bonded-neighbor 1−3 interactions.[43] For the divalent cations, the four terms representing the electrostatic interactions between the ion−Drude pair and the water oxygen−Drude pair were treated using eq 7.

2.3. Free Energy Calculations. The absolute hydration free energy of the ions was calculated and decomposed into three contributions following a free energy perturbation simulation protocol established previously[29,53]

$$\Delta G_{\mathrm{hydr}} = \Delta G_{\mathrm{rep}} + \Delta G_{\mathrm{disp}} + \Delta G_{\mathrm{elec}} \tag{8}$$

where ΔG_{rep} and ΔG_{disp} are the repulsive and attractive (dispersive) components, respectively, of the LJ interaction in eq 1.The electrostatic component of the hydration free energy is ΔG_{elec}. Each component of the total hydration free energy was computed from independent simulations in which an ion was placed in a periodic box containing 216 explicit SWM4-NDP water molecules. Long-range electrostatic interactions were computed using particle mesh Ewald summation.[54] A smooth real-space cutoff is applied between 10 and 12 Å with an Ewald splitting parameter of 0.34 $\mathrm{Å}^{-1}$, a grid spacing of ∼1.0 Å, and a sixth-order interpolation of the charge to the grid. The isothermal−isobaric ensemble was simulated using a Nosé−Hoover thermostat[55,56] and the modified Andersen−Hoover barostat of Martyna et al.[57] along with a 1 fs time step. The internal geometry of the SWM4-NDP water molecule was kept rigidly fixed using the SHAKE/Roll and RATTLE/Roll algorithm.[58,59] For each value of the thermodynamic coupling parameter, λ, after an initial equilibration of 200 ps, equilibrium properties were averaged over a 400 ps molecular dynamics simulation. For the dispersive and electrostatic components, λ took on values between $\lambda = 0$ and $\lambda = 1$ that were evenly spaced in increments of 0.1. For the repulsive term, λ took on the following values: 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0, 3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. The repulsive contribution, ΔG_{rep}, was computed using a soft-core scheme as described elsewhere[53] and unbiased using the weighted histogram analysis method (WHAM),[60] while ΔG_{disp} and ΔG_{elec} were computed using thermodynamic integration (TI). On the basis of multiple runs, the overall precision of the calculated absolute hydration free energies with the current protocol is on the order of 0.2 kcal/mol for monovalent ions and 0.5 kcal/mol for divalent ions.

In discussions of the hydration free energy of ionic species, one may consider the *real* physical value, which includes the contribution of the phase potential arising from crossing the physical air/water interface or the *intrinsic* bulk phase value, which is independent of the interfacial potential.[29,61,62] The relationship between *real* and *intrinsic* hydration free energies is defined as $\Delta G_{\mathrm{hydr}}^{\mathrm{real}} = \Delta G_{\mathrm{hydr}}^{\mathrm{intr}} + zF\Phi$, where F is the Faraday constant (23.06 kcal/mol/V) and Φ is the electrostatic Galvani potential at the liquid−vacum interface or the phase potential of the liquid relative to vacuum. It may be tempting to consider the intrinsic free energy as somehow reflecting more faithfully the true hydration free energy of the ion within the bulk phase because it is "disentangled" from the bias arising from the liquid−vacum interfacial potential. However, it should be noted that the actual value of Φ depends upon the convention to define the Galvani potential. One may construct an "internal" Galvani potential via a P-sum, where the potential at all points in space is the superposition of the total charge density from all the particles.[63,64] Alternatively, one may construct an "external" Galvani potential via a M-sum, where the potential from each water molecule contributes only to points in space that are outside their repulsive core.[63,64] Although the internal or external Galvani potentials can be defined mathematically without ambiguity by specifying which P- or M-sum convention has been used, they cannot be measured experimentally by a physical process.

To avoid any ambiguity, only *real* hydration free energies are considered throughout the present study. $\Delta G_{\mathrm{hydr}}^{\mathrm{real}}$ corresponds to the reversible thermodynamic work to move a single ion from vacuum to the interior of a pure water phase (i.e., across the physical liquid−vacuum interface). Our free energy calculations with periodic boundary conditions (PBC) and particle mesh Ewald are carried out within the P-sum convention, and the implicit reference phase potential of the liquid is the "internal" Galvani potential. Such free energy calculations yield $\Delta G_{\mathrm{hydr}}^{\mathrm{intr}}$, which then needs to be shifted by $z\Phi$ (calculated within the same P-sum convention) to yield the physically meaningful $\Delta G_{\mathrm{hydr}}^{\mathrm{real}}$. For our polarizable SWM4-NDP water model, the P-sum internal Galvani potential Φ is equal to −545 mV[35] (negative in the liquid phase relative to vacuum), giving rise to an energy shift of ∓12.6 kcal/mol for the monovalent cations/anions and twice that amount for the divalent species. Finally, for a direct comparison with published experimental tables, it is necessary to convert the results into the free energy of an ion going from an ideal gas at 1 atm to an idealized bulk solution at 1 M concentration. To account for the compression from a volume of 24.465 L/mol in the ideal gas to the 1 M solution, a small entropic contribution of $-k_{\mathrm{B}}T \ln (1/24.465) = 1.9$ kcal/mol must be added per ion.[29]

2.4. Target Data and Parametrization Strategy. *2.4.1. Ionic Charges and Polarizabilities.* For the alkali cations and halide anions, the values of the polarizabilities from the earlier parametrization compatible with the SWM-DP water model were used here without changes.[29] In Table 1 included among these polarizabilities are gas-phase values,

Table 1. Parameters for Alkali Cations, Halide Anions, and Divalent Cations with Negatively Charged Drude Oscillators

	q (lel)	α (Å^3)[a]	$-q_D$ (lel)	E_{min} (kcal/mol)	$R_{min}/2$ (Å)	Thole t_{iw}
Li^+	+1	0.032 (0.032)	0.310427	0.0300000	1.1000000	
Na^+	+1	0.157 (0.157)	0.687597	0.0315100	1.4616800	
K^+	+1	0.830 (0.830)	1.580968	0.1419265	1.6866521	
Rb^+	+1	1.370 (1.370)	2.031161	0.2730669	1.7855083	
Cs^+	+1	2.360 (2.360)	2.665877	0.2766036	2.0238218	
F^-	−1	1.786 (2.467)	2.319199	0.0026181	2.4622406	
Cl^-	−1	3.969 (5.482)	3.457187	0.0719737	2.4811139	
Br^-	−1	5.262 (7.268)	3.980713	0.0823440	2.6262883	
I^-	−1	7.439 (10.275)	4.733085	0.2084343	2.7579694	
Zn^{2+}	+2	0.420 (0.420)	1.124637	0.2500000	0.9349678	2.14773
Mg^{2+}	+2	0.075 (0.075)	0.475246	0.0500000	1.1264156	1.51567
Ca^{2+}	+2	0.490 (0.490)	1.214747	0.2100000	1.2708552	1.50877
Sr^{2+}	+2	0.870 (0.870)	1.618629	0.3400000	1.3059400	1.23792
Ba^{2+}	+2	1.560 (1.560)	2.167454	0.6000000	1.5717385	1.45869

[a] In parentheses are the ab initio estimates of the gas-phase polarizabilities from which the values for α are derived (see ref 65 for the alkali ions and ref 66 for the halide ions and divalent ions).

Table 2. Properties of Alkali Cations, Halide Anions, and Divalent Cations with Drude Polarizable Models

ion	U_{min}[a]	d_{min}[b]	ΔH[c]	r_{max}	g_{max}	r_{min}	g_{min}	N_c	D[d]	ΔG_{hydr}^{real}[e]	$\Delta\Delta G_{hydr}$[f]
Li^+	−35.9 (−35.2)	1.91 (1.87)	−35.6 (−34.0, −34.0)	2.02	12.50	2.56	0.00	4.0	1.30 (1.03)	−120.5	−24.2 (−23.8 to −26.2)
Na^+	−24.6 (−24.3)	2.25 (2.26)	−24.4 (−24.0, −25.0)	2.38	7.42	3.24	0.20	5.6	1.58 (1.33)	−96.3	−17.3 (−16.7 to −17.7)
K^+	−17.9 (−17.8)	2.62 (2.64)	−17.6 (−17.9, −18.1)	2.74	4.80	3.56	0.45	6.9	2.20 (1.98)	−78.6	−5.2 (−4.9 to −5.4)
Rb^+	−15.7 (−16.1)	2.79 (2.79)	−15.2 (−15.9, −16.0)	2.90	4.04	3.80	0.62	8.1	2.44 (2.07)	−73.7	−7.1 (−5.5 to −7.7)
Cs^+	−13.3 (−14.1)	3.05 (2.99)	−12.5 (−13.7, −)	3.16	3.25	4.10	0.75	9.7	2.56 (2.06)	−66.5	
F^-	−23.5 (−25.9)	2.53 (2.44)	−23.2 (−23.3, −23.3)	2.72	4.77	3.34	0.37	5.5	1.33 (1.48)	−108.0	−30.0 (−13.4 to −30.6)
Cl^-	−14.0 (−14.4)	3.09 (3.11)	−13.7 (−13.1, −14.4)	3.16	3.15	3.78	0.72	6.5	1.82 (2.03)	−78.4	−6.5 (−3.3 to −6.9)
Br^-	−12.4 (−12.7)	3.26 (3.26)	−11.9 (−12.6, −13.0)	3.28	2.70	3.96	0.75	6.8	1.85 (2.08)	−71.6	−8.5 (−7.7 to −11.1)
I^-	−10.2 (−10.6)	3.56 (3.50)	−9.7 (−10.2, −10.5)	3.50	2.28	4.16	0.90	7.1	2.02 (2.05)	−63.1	
Zn^{2+}	−100.0 (−96.3)	1.82 (1.86)	−99.4 (−103.1)	2.14	17.2	3.08	0.0	6.0	0.61 (0.70)	−460.2	−90.3 (−107.6)
Mg^{2+}	−89.4 (−77.9)	1.86 (1.93)	−89.0 (−81.8)	2.06	19.0	2.72	0.0	6.0	0.82 (0.71)	−447.2	−77.3 (−77.7 to −80.3)
Ca^{2+}	−55.6 (−54.9)	2.18 (2.25)	−55.1 (−56.5)	2.28	16.9	2.76	0.0	6.0	0.96 (0.79)	−369.9	−32.7 (−29.8 to −32.9)
Sr^{2+}	−45.2 (−40.6)	2.30 (2.52)	−44.7 (−)	2.42	11.9	3.20	0.0	7.2	0.96 (0.79)	−337.2	−27.2 (−27.9 to −31.1)
Ba^{2+}	−37.7 (−34.0)	2.56 (2.73)	−37.3 (−)	2.68	10.4	3.30	0.1	8.2	0.97 (0.85)	−310.0	

[a] The reference binding energy in parentheses U_{min} of monovalent monohydrates is taken from various sources:[29] Li^+,[73] Na^+,[74] K^+,[75] Rb^+,[74] Cs^+,[74] F^-,[76] Cl^-,[76] Br^-,[76] I^-.[76] The reference values for divalent ions were obtained with MP2 calculations with basis set 6-311++G(3df,3pd) for hydrogen, oxygen, Mg^{2+}, Ca^{2+}, and Zn^{2+} and LANL2DZ pseudopotential and associated basis set for Sr^{2+} and Zn^{2+}. The energies are in kcal/mol. [b] The reference distance in parentheses of the alkali cations is taken from various sources,[29] while those of the halide anions are taken from ref 76. The reference values for divalent cations were obtained from MP2 calculations (see footnote *a* of this table for details). The distances are in Ångstroms. [c] The reference enthalpies in parentheses ΔH for the monovalent monohydrates are taken from the experimental studies.[71,72] The reference enthalpies in parentheses ΔH for the divalent monohydrates are taken from the theoretical study.[87] The enthalpies are in kcal/mol. [d] The experimental self-diffusion coefficients are taken from the *Handbook of Chemistry and Physics*.[95] The self-diffusion coefficients are in 10^{-5} cm ^{2}/s. [e] The *real* hydration free energies ΔG_{hydr}^{real}, in kcal/mol, are obtained from the *intrinsic* free energies from periodic boundary condition (PBC) simulations corrected with the phase potential- and entropy-related contributions. [f] The experimental hydration free energy differences in kcal/mol for the monovalent ions are taken from refs 72 and 78−83, while those for the divalent ions are taken from refs 80, 82, and 83. The entry for Zn^{2+} is defined as $\Delta G_{hydr}^{real}(Zn^{2+}) - \Delta G_{hydr}^{real}(Ca^{2+})$.

reported by Mahan,[65] for the alkali cations and a set of "solvent-renormalized" values for the anion halides, with the bare gas-phase polarizability[66] scaled down by a factor of 0.724. Although the actual values vary, such reduced anion polarizabilities have also been noted in previous quantum mechanical studies;[67−70] see ref 29 for a more complete discussion. For the divalent cations, gas-phase values were again taken from Mahan:[65] 0.075 Å ^{3} for Mg^{2+}, 0.49 Å ^{3} for Ca^{2+}, 0.87 Å ^{3} for Sr^{2+}, 1.56 Å ^{3} for Ba^{2+}, and 0.42 Å ^{3} for Zn^{2+}.

2.4.2. Ionic Monohydrates. For the monovalent ions, the target monohydrate properties are the same as those used in a previous effort[29] and are summarized in Table 2. Namely, the target binding enthalpies, ΔH, for the alkali cations and halide anions are the gas-phase binding enthalpies measured by Džidić and Kebarle[71] and Tissandier et al.[72] In addition, the binding energies are taken from the ab initio computations.[73−76] A set of target ion−oxygen distances was established for the monovalent ions as described in detail by Lamoureux and Roux.[29] For the divalent cations, ab initio quantum chemical computations were carried out at the MP2 level using Gaussian03.[77] The details of the basis sets used and resulting target data are presented in Table 2. In these computations, the water geometry was fixed to that of the SWM4-NDP model while the ion−oxygen separation was varied.

2.4.3. Absolute Hydration Free Energies. The experimentally determined absolute hydration free energies of ions at infinite dilution are a central piece of information to optimize the model.[72,78−83] However, as noted previously,[29] there is a large spread in the reported values for the alkali and halide ions (see Figure 2 of ref 29). The same is true for the divalent cations. Considerations based on interfacial potentials may help explain the origin of some of those discrepancies, although it is important to realize that the overall offset of the absolute hydration free energy scale measured experimentally remains undetermined. The most reliable target data that can be extracted from the experimental measurements are the

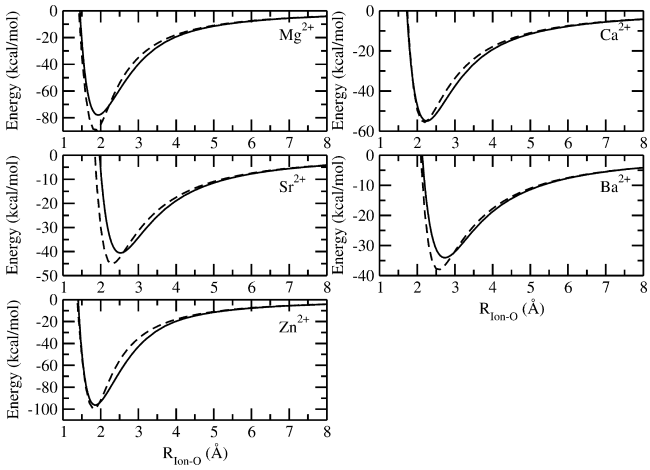


Figure 2. Binding energies of the divalent monohydrates as a function of the distance between the ion and the oxygen atoms: (solid line) ab initio results; (dashed line) Drude. See Table 2 footnotes for the theory and basis set used to obtain the reference data.

relative hydration free energies between different ions of identical charge and the total solvation free energies of neutral salts ($\Delta G_{hydr}(AB_n) = \Delta G_{hydr}(A) + n\,\Delta G_{hydr}(B)$). Notwithstanding those issues, it should be noted there are also some quantitative differences among the various experimental values. For example, Tissandier et al.[72] reported the total hydration free energy of KCl to be −156.8 kcal/mol, while Randles[78] reported −151.3 kcal/mol. The values from Tissandier,[72] Schmid,[83] and Klots[81] are quite close for the monovalent ions. During the parametrization of monovalent ions, we principally aimed to reproduce the target data generated from the measurements of Tissandier.[72] For the divalent ions, the target data was taken from the measurements of Schmid,[83] except in the case of Zn^{2+} where our model falls between the data from Schmid[83] and Gomer and Tryson.[80]

2.4.4. Optimization Procedure. The parameters for the ion−water oxygen interaction are constructed via the Lorentz−Berthelot combination rule,[46] $E^{iw}_{min} = (E^{ion}_{min}E^{O}_{min})^{1/2}$ and $R^{iw}_{min} = (R^{ion}_{min} + R^{O}_{min})/2$. The adjustable parameters for the monatomic ions within the classical Drude oscillator scheme are the LJ parameters of the ion, R^{ion}_{min} and E^{ion}_{min}. Rather than exploring the LJ parameters of the ions directly, it is more convenient to explore the space of monohydrate interaction energies and minimum-energy ion−oxygen distances $\{d_{min}, U_{min}\}$,[29] which can be compared to the available ab initio or experimental data. Furthermore, quadratic response functions are fitted to the data from explicit computations, defined by coordinates in $\{d_{min}, U_{min}\}$, to interpolate predicted properties (e.g., hydration free energy) between simulated models[29,35]

$$\Delta\bar{G}_{hydr}(d_{min}, U_{min}) = a_0 + a_1 d_{min} + a_2 U_{min} + a_3 d^2_{min} + a_4 d_{min} U_{min} + a_5 U^2_{min} \quad (9)$$

A set of polarizable models for the monovalent ions was thus constructed by determining the LJ parameters spanning a grid in the $\{U_{min}, d_{min}\}$ coordinates and simultaneously fitting all ions using the interpolated polynomials. For divalent cations that use the Thole screening as described by eq 7, there are three parameters to fit per ion: the LJ radius and well depth of the ion and the ion−water Thole screening factor. To simplify the parametrization process and ensure physically reasonable dispersive interactions, it proved useful to first assign the LJ well depth by making use of the familiar London dispersion formula $\sim C^{ij}_6/r^6$ for the leading order dispersion coefficient

$$C^{iw}_6 = \frac{3}{2}\left(\frac{E^{i}_I E^{w}_I}{E^{w}_I + E^{w}_I}\right)\alpha_i\alpha_w \quad (10)$$

where E^{i}_I, E^{w}_I, α_i, and α_w are the ionization energy and polarizabilty of the ion and water, respectively. Converting C^{iw}_6 into the attractive coefficient of the LJ interaction, $2E^{iw}_{min}(R^{iw}_{min})^6$, yields an estimate of E^{iw}_{min}. The resulting values are 0.05 for Mg^{2+}, 0.21 for Ca^{2+}, 0.34 for Sr^{2+}, 0.60 for Ba^{2+}, and 0.25 for Zn^{2+} (all in kcal/mol). Once the LJ well depth is assigned, a grid in $\{d_{min}, U_{min}\}$ was generated for each divalent ion by varying the LJ radius and Thole parameter for each ion. The parametrization of the complete set of ion models is fairly consistent and robust. As observed previously,[29,84] alternative sets of parameters can be found with small variations in the ion monohydrate energies (±0.25 kcal/mol) or in the ion−water distances (±0.1 Å), resulting in small shifts in the absolute hydration free energies (±3 kcal/mol). At this stage, the parametrization procedure is aimed at designing accurate models for the infinite dilution limit. Solutions at high concentrations, where ion−ion interactions become important, will be examined in a later stage. The final fitted values are summarized in Table 1.

3. Results and Discussion

The previously developed K^+ model was shown to capture the essential properties in both the gas phase and aqueous solution.[84] The parameters for the polarizable Drude ion models were optimized to be consistent with bulk hydration free energies of the neutral salts while yielding accurate energies and geometries for the ionic monohydrates. An important advantage of reproducing the bulk hydration free energies of neutral salts with our model ions is that these data contain no undetermined offset constant, as do the absolute hydration free energies of individual ions.[29] Achieving an accurate description for the ionic monohydrates becomes critical, therefore, to "lock down" the absolute scale for the bulk hydration free energies. As shown in Table 2, this procedure defines a "solvation-consistent" scale for the absolute hydration free energies of ionic species despite the uncertainties in the experimental data. Overall, the properties of bulk solvation given in Table 3 and small gas-phase clusters given in Table 4 are in excellent agreement with experiments.

3.1. Instabilities and Overpolarization. The polarizable Drude models based on the simple scheme presented by eq 1 yield numerically stable simulations for all the monovalent cations. As expected, based upon substitution into eq 4, no numerical instabilities based on overpolarization were present. In contrast, eq 4 can be used to predict overpolarization catastrophes that occur for halide anions (Br^- and I^-) when

Table 3. Hydration Free Energy of Neutral Salts (in kcal/mol)[a]

	Li^+	Na^+	K^+	Rb^+	Cs^+
F^-	−228.5 (−229.0)	−204.3 (−203.8)	−186.6 (−186.6)	−181.7 (−181.2)	−174.5 (−173.8)
Cl^-	−198.9 (−199.3)	−174.7 (−174.0)	−157.0 (−156.8)	−152.1 (−151.4)	−144.9 (−144.0)
Br^-	−192.1 (−192.9)	−167.9 (−167.6)	−150.2 (−150.4)	−145.3 (−145.0)	−138.1 (−137.6)
I^-	−183.6 (−183.9)	−159.4 (−158.7)	−141.7 (−141.5)	−136.8 (−136.1)	−129.6 (−128.7)

	Zn^{2+}	Mg^{2+}	Ca^{2+}	Sr^{2+}	Ba^{2+}
F^-	−676.2 (−696.8)	−663.2	−585.9 (−589.4)	−553.2	−526.0 (−528.1)
Cl^-	−617.0 (−637.2)	−604.0 (−607.0)	−526.7 (−529.9)	−494.0 (−499.0)	−466.8 (−468.6)
Br^-	−603.4 (−624.4)	−590.4 (−594.3)	−513.1 (−517.0)	−480.4 (−486.2)	−453.2 (−455.7)
I^-	−586.4 (−606.5)	−573.4 (−576.4)	−496.1 (−499.1)	−463.4 (−468.3)	−436.2 (−437.8)

[a] The experimental solvation free energies reported for alkali halide salts by Tissandier et al.[72] and Klots[81] (Cs^+) are listed in parentheses for comparison. The experimental solvation free energies for salts of divalent cations were derived from the free energies of formation of the salt minus that of the gas-phase ions (in kJ/mol): $\Delta G^\circ_{aq}(A^{n+} + nB^-) = \Delta G^\circ_f(AB_n) - \Delta G^\circ_f(A^{n+}(g)) - n\Delta G^\circ_f(B^-(g))$, with $\Delta G^\circ_f(A^{n+}(g)) = \Delta H^\circ_f(A^{n+}(g)) - T\Delta S^\circ_f(A^{n+}(g))$ and $\Delta S^\circ_f(A^{n+}(g)) = 108.8555 + 12.4716 \ln(M) - \nu^* S(A^*) \pm nS(e^-(g))$, where M is the molar mass in grams, ν^* is 1/2 for anions and 1 for cations, $S(A^*)$ is the element's absolute entropy in the standard state, and $S(e^-(g))$ is the absolute entropy of the gaseous electron (see ref 72 for details). The experimental thermodynamic data are taken from the *Handbook of Chemistry and Physics*[95] and the 1982 NBS tables.[85]

Table 4. Solvation Enthalpies of Ion M in Water Clusters $(H_2O)_n$ (in kcal/mol)

		solvation enthalpy					
ions, M	model/exp	1	2	3	4	5	6
Li^+	Drude	−35.6	−65.3 (−29.7)	−87.0 (−21.7)	−103.0 (−16.0)	−114.2 (−11.2)	−125.3 (−11.1)
	exp[72]	−34.0	−59.8 (−25.8)	−80.5 (−20.7)	−96.9 (−16.4)	−110.8 (−13.9)	−122.9 (−12.1)
	exp[71]	−34.0	−59.8 (−25.8)	−80.5 (−20.7)	−96.9 (−16.4)	−110.8 (−13.9)	−122.9 (−12.1)
Na^+	Drude	−24.4	−45.9 (−21.5)	−63.4 (−17.5)	−77.2 (−13.8)	−87.0 (−9.8)	−95.1 (−8.1)
	exp[72]	−25.0	−44.8 (−19.8)	−60.2 (−15.4)	−73.4 (−13.2)	−84.9 (−11.5)	−95.6 (−10.7)
	exp[71]	−24.0	−43.8 (−19.8)	−59.6 (−15.8)	−73.4 (−13.8)	−85.7 (−12.3)	−96.4 (−10.7)
K^+	Drude	−17.6	−33.2 (−15.6)	−47.0 (−13.8)	−58.6 (−11.6)	−67.7 (−9.1)	−76.0 (−8.3)
	exp[72]	−18.1	−34.2 (−16.1)	−47.4 (−13.2)	−59.2 (−11.8)	−69.9 (−10.7)	−79.9 (−10.0)
	exp[71]	−17.9	−34.0 (−16.1)	−47.2 (−13.2)	−59.0 (−11.8)	−69.7 (−10.7)	−79.7 (−10.0)
Rb^+	Drude	−15.3	−29.0 (−13.7)	−41.2 (−12.2)	−51.9 (−10.7)	−60.3 (−8.4)	−68.5 (−8.2)
	exp[72]	−16.0	−29.6 (−13.6)	−41.8 (−12.2)	−53.0 (−11.2)	−63.5 (−10.5)	
	exp[71]	−15.9	−29.5 (−13.6)	−41.7 (−12.2)	−52.9 (−11.2)	−63.4 (−10.5)	
Cs^+	Drude	−12.4	−24.3 (−11.9)	−34.7 (−10.3)	−44.0 (−9.3)	−52.0 (−8.0)	−64.0 (−12.0)
	exp[71]	−13.7	−26.2 (−12.5)	−37.4 (−11.2)	−48.0 (−10.6)		
F^-	Drude	−23.2	−45.4 (−22.2)	−66.9 (−21.5)	−80.2 (−13.3)	−90.7 (−10.5)	−103.1 (−12.4)
	exp[72]	−23.3	−41.2 (−17.9)	−55.7 (−14.5)	−69.4 (−13.7)	−82.2 (−12.8)	−93.1 (−10.9)
Cl^-	Drude	−13.7	−27.6 (−13.9)	−43.9 (−16.3)	−57.4 (−13.5)	−64.4 (−7.0)	−71.7 (−7.3)
	exp[72]	−14.4	−27.2 (−12.8)	−38.9 (−11.7)	−49.8 (−10.9)	−59.3 (−9.5)	−68.1 (−8.8)
Br^-	Drude	−11.9	−24.6 (−12.7)	−40.4 (−15.8)	−53.2 (−12.8)	−59.6 (−6.4)	−68.1 (−8.5)
	exp[72]	−13.0	−25.0 (−12.0)	−36.4 (−11.4)	−47.4 (−11.0)	−58.2 (−10.8)	−68.5 (−10.3)
I^-	Drude	−9.7	−20.9 (−11.2)	−35.6 (−14.7)	−42.4 (−6.8)	−50.7 (−8.4)	−59.6 (−8.9)
	exp[72]	−10.5	−20.2 (−9.7)	−29.5 (−9.3)	−38.7 (−9.2)	−47.7 (−9.0)	
Zn^{2+}	Drude	−99.4	−176.9 (−77.5)	−237.2 (−60.3)	−284.4 (−47.2)	−306.2 (−21.8)	−329.0 (−22.8)
	theor or exp[87]	−103.1	−191.1 (−88.0)	−246.9 (−55.8)	−289.8 (−42.9)	−314.8 (−25.0)	−339.0 (−24.2)
Mg^{2+}	Drude	−89.0	−164.6 (−75.6)	−223.3 (−58.7)	−269.8 (−46.5)	−293.0 (−23.2)	−316.5 (−23.5)
	theor or exp[87]	−81.8	−153.8 (−72.0)	−210.4 (−56.6)	−254.3 (−43.9)	−282.3 (−28.0)	−306.9 (−24.6)
Ca^{2+}	Drude	−55.1	−105.2 (−50.1)	−149.5 (−44.2)	−188.1 (−38.6)	−217.8 (−29.7)	−245.3 (−27.5)
	theor or exp[87]	−56.5	−105.1 (−48.5)	−148.0 (−42.9)	−183.6 (−35.6)	−211.3 (−27.7)	−236.6 (−25.3)
Sr^{2+}	Drude	−44.7	−86.2 (−41.5)	−124.0 (−37.8)	−158.4 (−34.4)	−187.0 (−28.6)	−213.6 (−26.6)
	exp[86]						− (−22.7)
Ba^{2+}	Drude	−37.3	−72.0 (−34.7)	−103.7 (−31.7)	−133.3 (−29.6)	−158.7 (−25.4)	−182.6 (−23.9)
	exp[86]					− (−23.9)	− (−19.8)

[a] Numbers in parentheses are for the reaction $M(H_2O)_{n-1} + H_2O \rightleftharpoons M(H_2O)_n$.

ion−proton separations are within physically realistic distances. Such instabilities are not present in our previously published set of alkali−halide ionic models[29] because the positive Drude particle and proton do not attract one another. Introducing an anharmonic restoring force to the anionic Drude oscillator removes those instabilities by preventing excessively large excursions of the Drude particle away from the ion. Importantly, as shown in Tables 2−4, this construction did not affect our ability to adjust the parameters of the model to accurately fit the target data. In particular, the properties of the small cluster hydrates can be reproduced with the simple anharmonic restoring spring.

A more serious overpolarization problem is encountered with the divalent cations. We first tried the two simple schemes to solve this issue, namely, adding an anharmonic restoring force to the Drude oscillator or adding a LJ core repulsion between the oxygen-tethered Drude particle and the divalent cations. Experimenting with various parametrizations showed that while introducing an anharmonic restoring force or a LJ core can prevent the numerical

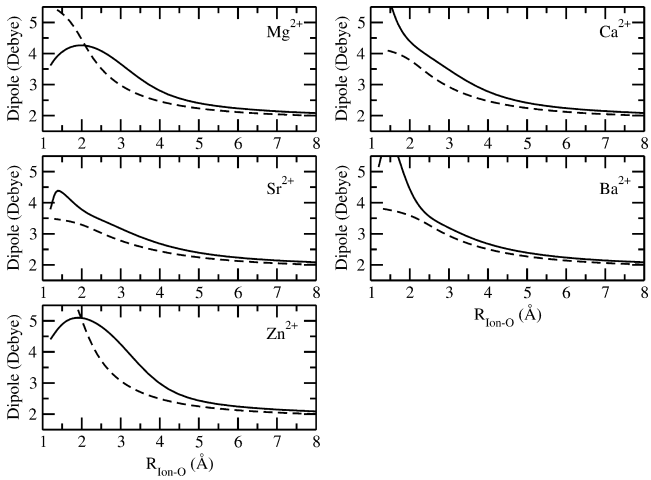


Figure 3. Total dipole moments of the divalent monohydrates as a function of the distance between the ion and the oxygen atoms. The ions are located at the origin to compute the dipole moment. See Table 2 for details.

instabilities, it remained difficult to get reasonable monohydrate energies as the large Coulombic repulsion between the bare positive charge carried by the oxygen nucleus and the ion remains unshielded. This Coulombic interaction gives rise to large repulsion energies that must be compensated in order to have reasonable monohydrate energies and geometries with the divalent cations. Ultimately, it was found that introducing a Thole-type screening function[49] for the Coulombic electrostatic for the interactions between the divalent ion and the induced dipole component of the water molecules led to the best models. A similar approach was used to construct models of divalent ions consistent with the AMOEBA polarizable force field.[30,52] Comparisons between the Drude model with Thole-like damping and the ab initio calculations are shown in Figures 2 and 3. The Thole screening scheme given by eq 7 effectively removes the singularity problems illustrated in Figure 1 and yields divalent monohydrates with reasonable binding energies and total dipole moments.

3.2. Bulk Hydration Free Energies. The total hydration free energies of neutral salts are shown in Table 3 and Figures 4 and 5 together with available experimental data. Although the relative hydration free energy difference between different ions in the same series and the total hydration free energy of a neutral salt can generally be measured very accurately, we note that there are noticeable differences among the experimental data from difference sources[72,78−83,85] (see Supporting Information Tables S1 and S2). Overall, the current Drude models agree very well with the experimental data, with the monovalent neutral salts closer to the experimental estimations by Tissandier et al.,[72] Klots,[81] and Schmid et al.[83] (Figure 4) and the divalent neutral salts closer to the experimental estimations by Schmid et al.[83] and Gomer and Tryson[80] (Figure 5).

The hydration free energy differences of ions in the same series are well reproduced with some compromise of the monohydrate properties for smaller ions. These present models define an absolute hydration free energy scale in which the *real* hydration free energy of the proton would be

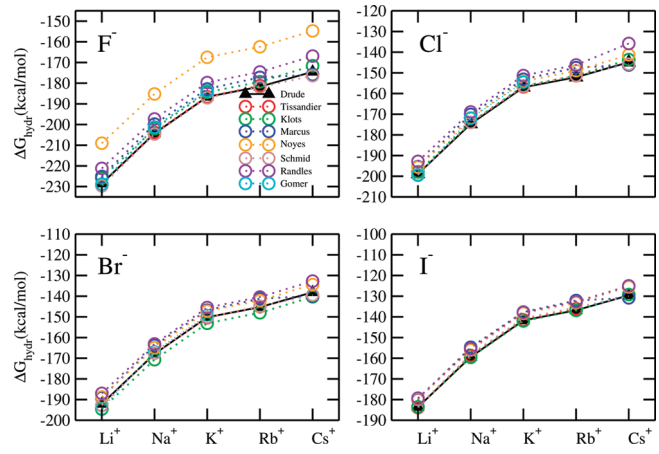


Figure 4. Hydration free energies of salts with monovalent cations: Tissandier,[72] Klots,[81] Marcus,[82] Noyes,[79] Schmid,[83] Randles,[78] and Gomer.[80]

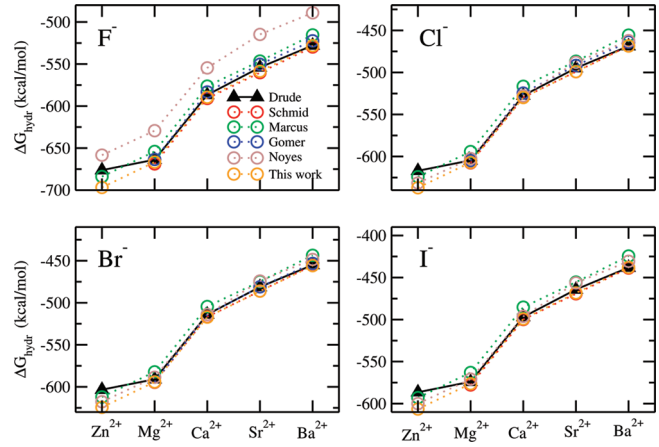


Figure 5. Hydration free energies of salts with divalent cations: Schmid,[83] Marcus,[82] Gomer,[80] Noyes,[79] and this work. See the caption of Table 3 for details. The entries for MgF_2 and SrF_2 in this work were derived based on hydration free energy differences between F^- and Cl^- and the total hydration free energies of salts $MgCl_2$ and $SrCl_2$.

−258.8 kcal/mol. This can be compared with the experimental estimates of −264 kcal/mol by Tissandier et al.,[72] −249.5 kcal/mol by Schmid et al.,[83] and −253.2 kcal/mol by Gomer and Tryson.[80] It is noteworthy that the current value of −258.8 kcal/mol differs by less than 2% from the experimental estimates obtained by Tissandier et al.[72] This suggests that their analysis, which consisted of monitoring the stepwise changes in ion−cluster free energies, was targeting the absolute *real* free energy of the proton.[72]

3.3. Enthalpies of Ionic Small Hydrates. Structural and thermodynamic properties of small clusters offer a rigorous test of the accuracy of the models. As opposed to bulk liquid hydration where induced polarization over different moieties may lead to some compensation of errors, the microscopic interactions are displayed nakedly in small clusters, and possible limitations of the models become readily apparent. Furthermore, because the polarizable models are constructed with the intended purpose of responding accurately to different electrostatic environments, the ability to reproduce the properties of small ionic hydrates is very important. Finally, the properties of the ionic monohydrates are used

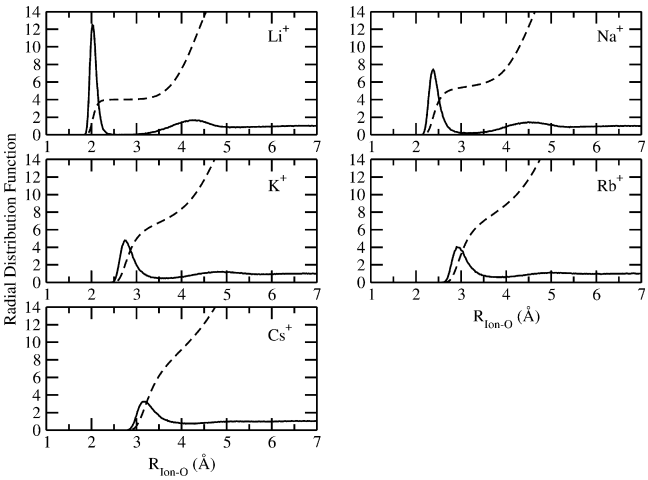


Figure 6. Radial hydration structure for the alkali cations. Radial distribution functions $g(R_{Ion-O})$ functions are shown by the solid line, and the coordination numbers $N(R_{Ion-O})$ are shown by the dashed line.

to set the scale for the absolute hydration free energies in bulk solution. The calculated ionic monohydrate energies and geometries are reported in Table 2. The results generally agree well with the target data.[73−76] As previously noted,[29] the limitations of the models are most apparent for smaller ions. This trend is observed for the divalent cations as well. The less satisfactory performance of the divalent cations is very likely due to the neglect of the important charge-transfer effect, which is absent from the current energy function. The enthalpies of hydration for an ion solvated in a small water cluster (1−6) compare favorably with literature values with the monovalent cations having the best agreement[71,72,86,87] (Table 4). In general, the monovalent cations reproduce the experimentally measured properties of ionic hydrate clusters slightly better than do the halide anions. The most likely explanation for this is that the conformational spaces for water molecules in the anion−water clusters are far more complicated than those for the cation−water clusters due to the formation of hydrogen bonding between water molecules coordinating the anion. The hydration enthalpies for the divalent cations have larger deviation from the experimental or ab initio data, consistent with the monohydrate properties.

3.4. Coordination Structure in Bulk Solution. Radial distribution functions (RDF) averaged from 10 simulations each of 200 ps for ion−oxygen contacts, $g(R_{Ion-O})$, are presented in Figures 6, 7, and 8 for the alkali cations, halide anions, and divalent ions, respectively. The coordination numbers for each ion, $N(R_{Ion-O})$, defined as the integral of $g(R_{Ion-O})$ from the origin out to the first minimum, r_{min}, are presented in Table 2. The RDFs and coordination numbers for the monovalent ions are very similar to those reported from the previous models (Figures 6 and 7) based upon the (positive Drude) SWM4-DP water model.[29] Here, we briefly discuss the main findings.

In accordance with the most recent computational studies, the current model predicts that the Li^+ ion is 4-coordinated within the experimental range obtained from neutron and X-ray diffraction (see refs 88−90 for a survey). After the first peak, the lithium−oxygen RDF remains very small over a large interval extending from 2.4 to 3.1 Å. This result is

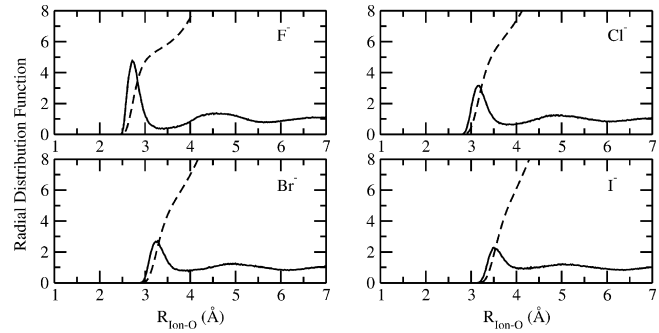


Figure 7. Radial hydration structure for the halide anions. Radial distribution functions $g(R_{Ion-O})$ are shown by the solid line, and the coordination numbers $N(R_{Ion-O})$ are shown by the dashed line.

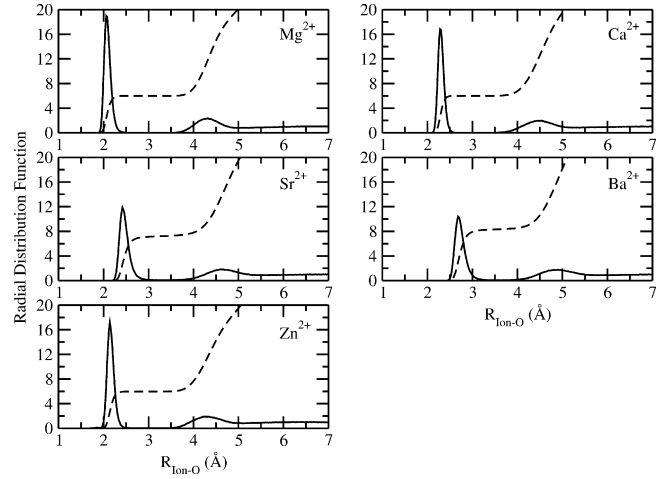


Figure 8. Radial hydration structure for the divalent cations. Radial distribution functions $g(R_{Ion-O})$ functions are shown by the solid line, and the coordination numbers $N(R_{Ion-O})$ are shown by the dashed line.

consistent with Car−Parrinello molecular dynamics (CPMD) simulation.[91] The strong separation between the first and second peak in the RDF implies that the coordination number of Li^+ can be defined unambiguously. However, for Na^+ and K^+, a more populated region is observed between the first and second peaks in the RDFs. As a result, a broad distribution of coordination numbers within the first solvation shell was observed.[22,84] As dissussed in refs 29, 84, and 90, both experimental studies and CPMD-type simulations have predicted a wide range of coordination numbers for Na^+ and K^+ and the current models fall well within the reported range. In addition, the coordination number distributions of Na^+ and K^+ overlap substantially with one another, indicating that the coordination number alone is not a dominant factor in describing the thermodynamics of ion solvation.[22]

For the divalent ions, the first peak in the RDF is generally sharper than those of the monovalent ions, indicating a highly ordered water structure around the ions (Figure 8). In addition, a clear separation exists between the first and the second coordination shells, as indicated by the small value of the first minimum in the RDFs. The current Drude model for Ca^{2+} gives a coordination number of 6, smaller than the value of 7.3 based on the AMOEBA force field.[30] As a comparison, the coordination numbers of Ca^{2+} were esti-

Table 5. Effective Self-Diffusion Coefficients[a] of the Neutral Salts in 10^{-5} cm^2/s (the experimentally derived values are in parentheses)[95]

	Li^+	Na^+	K^+	Rb^+	Cs^+
F^-	1.32 (1.25)	1.45 (1.40)	1.77 (1.73)	1.89 (1.78)	1.95 (1.77)
Cl^-	1.56 (1.53)	1.70 (1.68)	2.01 (2.01)	2.13 (2.05)	2.19 (2.05)
Br^-	1.58 (1.56)	1.72 (1.71)	2.02 (2.03)	2.15 (2.08)	2.21 (2.07)
I^-	1.66 (1.54)	1.80 (1.69)	2.11 (2.01)	2.23 (2.06)	2.29 (2.06)

	Zn^{2+}	Mg^{2+}	Ca^{2+}	Sr^{2+}	Ba^{2+}
F^-	0.85 (0.96)	0.99 (0.97)	1.08 (1.02)	1.08 (1.02)	1.09 (1.06)
Cl^-	1.01 (1.14)	1.15 (1.15)	1.25 (1.20)	1.25 (1.20)	1.25 (1.25)
Br^-	1.02 (1.16)	1.16 (1.17)	1.26 (1.22)	1.26 (1.22)	1.26 (1.26)
I^-	1.08 (1.15)	1.22 (1.16)	1.31 (1.21)	1.31 (1.21)	1.32 (1.25)

[a] The effective self-diffusion coefficients for the neutral salts are defined as $D_{eff} = \sum_i z_i^2 D_i / \sum_i z_i^2$.

mated to lie between 5.5 and 10 based on various experimental techniques[92] and between 6 and 7 according to the most recent CPMD simulations.[92] For Mg^{2+}, the Drude model gives a coordination number of exactly 6, consistent with both the experimental[93] and the ab initio[94] studies. For Zn^{2+}, the coordination number is estimated to be 6, which is within the experimentally estimated range from 5.3 to 6.6.[88] Overall, the coordination structure for the ion models is in excellent agreement with experiment, given that no direct adjustments were made to reproduce those features during the parametrization.

3.5. Self-Diffusion Coefficients. The ability of the polarizable force field to reproduce the dynamic transport properties of ions in aqueous solutions is a critical test of the accuracy of computational models. This ability is best characterized by considering the diffusion of ions in aqueous solutions. The self-diffusion coefficients for the ion models at infinite dilution were computed from the mean-square displacement correlation function. Because there is only a single ion in the atomic system, relatively long simulations are required to obtain converged self-diffusion coefficients. To improve convergence, 10 simulations of 200 ps each were averaged for each ion. The results are presented in Tables 2 and 5. Additional simulations indicate that the current estimates from MD are not overly sensitive to finite-size effects within PBC.

The agreement between the Drude model and the available experimental values[95] is relatively good. Interestingly, the calculated values appear to be slightly overestimated for the cations while they are underestimated for the anions. This intriguing trend, with the cations diffusing faster and the anions slower than the experimental estimates, has also been noted in a recent study with nonpolarizable models.[17] Limitations in the ions and/or solvent models could perhaps explain these discrepancies. However, it is worth recalling that the polarizable SWM4-NPD water model matches the experimental value of the self-diffusion of water exactly, and the shear viscosity is satisfactorily reproduced as well.[35] Alternatively, there is also the possibility that the experimental estimates for the single ions are at fault for the following reason. Experimentally, it is nearly impossible to measure the self-diffusion coefficient of isolated ions in solution. Most of the available values for individual cations and anions have been deduced from conductivity measurements of electrolyte aqueous solutions by invoking additional hypotheses.[95] In the limit of very low concentration $[c]$, the conductivity σ of an electrolyte solution varies as $[c](e^2/k_BT)\sum_i z_i^2 D_i$. In practice, all the ionic species of a neutral salt contribute to the experimentally observed σ, suggesting that a more meaningful comparison is achieved by considering the effective self-diffusion coefficients for neutral salts, $D_{eff} = \sum_i z_i^2 D_i / \sum_i z_i^2$, corresponding to the low concentration limit of the whole salt conductivity after removing the dependence on the trivial prefactor $[c](e^2/k_BT)\sum_i z_i^2$.

In Table 5 the calculated and experimental values of D_{eff} for 40 neutral salts are listed. The agreement between the experimental data and the calculations is striking. With the exclusion of salts involving the smallest anion, F^-, the calculated values for the alkali−halides are almost perfect, while the salts with the divalent cations are all less than a few percent off. The remarkable agreement shown in Table 5 strongly argues in favor of the importance of considering the experimental transport properties of neutral salts rather than reported values for the self-diffusion coefficients of individual ions.

3.6. Comparison with Other Models. Over the years, a number of nonpolarizable[13−18,96] and polarizable[28−30,84] models of the monovalent ion series have been developed and parametrized to reproduce different sets of target data. Previous nonpolarizable models have been parametrized primarily on the basis of the absolute hydration free energies of single ions.[13−15,18,96] Recently, Joung and Cheatham pushed the limit of nonpolarizable force fields by enforcing a balanced description of crystal and solution properties.[16,17] Induced polarization is not a dominant factor for the hydration of a large cation like K^+, and satisfactory results can be obtained with nonpolarizable models.[84] Nevertheless, adopting nonpolarizable models often leads to some compromises concerning the gas-phase properties (e.g., the monohydrate properties)[15] or the dynamic propeties (e.g., the self-diffusion constant).[17] Furthermore, the nonpolarizable models are calibrated for bulk solutions, and their validity in different environments is uncertain. In contrast, the polarizable force fields improve the transferability by capturing the electronic polarization in different environments.[28−30,84] The AMOEBA models, an extension of the induced dipole model, developed by Ponder and co-workers,[28,30] accurately reproduce the various gas-phase and condensed-phase properties of ions, with more complex terms for electrostatics. One important difference of the polarizable models developed here compared to AMOEBA, where multipolar electrostatic terms are required, is that all the interactions are charge−charge Coulombic interactions and the functional form of the potential energy terms can be implemented in standard MD codes without major changes.

Another important difference concerns the parametrization strategy. Here, the parametrization sought to reproduce the well-established experimental data (i.e., the total hydration free energy of neutral salts as well as the relative free energy difference between different ions in the same series). By virtue of this procedure, the monohydrate properties served to lock down the absolute free energy scale. The parametri-

zation of the ion models in AMOEBA followed a somewhat similar route, focusing mainly on the monohydrate properties and the neutral salts.[28,30] This is in contrast with other efforts aimed at fitting the actual single-ion hydration free energies presented by a given experimental scale.[13−18,96] As noted previously,[29] we consider this to be problematic because the absolute hydration free energies measured experimentally depend on an undetermined phase potential, which is reflected in the large spread in the reported values. Coincidentally, the final models presented here happen to be close to the single-ion hydration free energies presented by Noyes for the monovalent cations,[79] although this scale was not imposed during parametrization. However, the results from the final models differ markedly from the single-ion hydration free energies presented by several of the experimental scales, and if those values were imposed as target data, then the unavoidable consequence would be a deterioration of the monohydrate properties. These considerations highlight the pitfalls with using the absolute single-ion hydration free energy reported from experiments as target data for force field parametrization.

4. Conclusion

In the current study, parameters for both cations and anions (Table 1) were developed for the polarizable force field based on classical Drude oscillator in conjunction with the SWM4-NDP water model. The parameters for the ions have been derived based on the gas-phase monohydrate properties and the hydration free energies in the bulk phase. The various gas-phase and condensed-phase properties (Tables 2−4) are in reasonable agreement with the available experimental data and ab initio calculations. The dynamic transport properties of the models are in excellent agreement with experimental estimates based on electrolyte conductivity (Table 5). Overall, the set of models developed here should provide an important tool for accurate studies of a wide range of biological and physical processes involving ions in aqueous solutions.

The models for divalent cations from the present work are somewhat less accurate than those of the monovalent ions. The results from ab initio quantum chemistry calculations and the hydration free energy are not perfectly reproduced. This could be in part due to the neglect of charge-transfer effects and to the breakdown of linear response.[97] As shown previously, the deviations from linear polarization response become quite severe with a perturbation by a charge of $+2.0e$.[97] In this case, it was necessary to treat the overpolarization that occurs when the ion is in direct contact with a water molecule. Analysis of the functional form of the potential function showed that an overpolarization catastrophe is unavoidable when the distance between a polarizable water and the divalent cation is less than 1.97 Å. By contrast, monovalent cations require no special treatment. Several strategies to avoid overpolarization for the divalent ions were considered. The best approach turned out to be introduction of a Thole-like screening function to modulate the Coulomb electrostatic interactions between the cation and the water molecule. Careful parametrization of this functional form against the results from quantum mechanical calculations led to reasonably accurate models. A similar approach was used to model Mg^{2+} and Ca^{2+} in the context of the AMOEBA polarizable force field.[30,52]

The present effort at parametrizing a complete set of ion models underlies the difficulties that arise when considering single-ion bulk properties deduced from experiments as objective target data. Experimental measurement of ions in bulk liquid phases are performed on globally neutral systems, and deriving single-ion properties requires additional nonthermodynamic hypotheses. The most meaningful comparisons are made with data directly obtained from neutral salts. For example, while the models do not match the single-ion diffusion coefficients deduced from electrolyte conductivity measurements, the effective diffusion coefficients for the neutral salts are in remarkably good agreement with experiments (Table 5). Such issues become particularly accute with respect to single-ion hydration free energies. Experimental estimates often rely on the free energy associated with the solvation of H^+ as a reference. However, published experimental scales for the alkali−halides solvation free energy can shift up and down, depending upon the chosen reference. Some of the experimental values may appear to be inconsistent with one another, which increases the difficulties in developing an accurate force field. These problems are further compounded by the fact that absolute free energies of charged species are affected by the air−liquid interfacial potential. While the solvation free energies of neutral salts reported in experimental tables are unambiguous and independent of the interfacial potential, absolute hydration free energies of ions are somewhat ambiguous and cannot be readily utilized as the target data for parametrization. Preserving complete consistency of the solvation free energies of all ions has been a critical aspect of the present effort. To extract a reliable set of target data for the absolute hydration free energy as required to parametrize our models, we previously adopted a strategy based on a "solvation-consistent" scale.[29] The consequences of this scale are well illustrated by considering the K^+ model. The parameters of K^+ have been adjusted to fit the properties of small cluster hydrates in the gas phase (Table 4). Once used in bulk-phase simulations, this model yields a (real) hydration free energy of −78.6 kcal/mol (Table 2). Further validating the K^+ model was the level of agreement of K^+−water oxygen radial distributions for the Drude model with results from both experimental and ab initio Born−Oppenheimer and Car−Parrinello MD simulations.[84] Therefore, the absolute free energy scale has been "locked down" by using the K^+ model, which then serves as a central reference for other ionic solutes. This suggests that the parametrization of additional charged moieties should be developed to be in accord with the present solvation-consistent scale.

Finally, it is worth pointing out that at this stage the models have been optimized to be accurate in the infinite dilution limit. The problem associated with electrolyte solutions at high concentrations, where issues of solubility and ion−ion interactions are important, will be examined in the near future. In particular, the models will be tested by calculating

the osmotic pressure of concentrated salt solutions using a new method.[98]

Acknowledgment. H.Y. is supported by a postdoctoral fellowship from the Joint Theory Institute of the University of Chicago and Argonne Laboratory. This work was funded in part by NIH grants GM072558 and GM051501.

Supporting Information Available: Comparison of the induced dipole moments of Cl^- and Br^- as a function of external electrostatic field strength, and compilation of the experimentally derived hydration free energy differences between different ions. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hille, B. *Ion channels of excitable membranes*, 3rd ed.; Sinauer Associates: Sunderland, MA, 2001.

(2) Page, M. J.; Di Cera, E. *Physiol. Rev.* **2006**, *86*, 1049–1092.

(3) Dudev, T.; Lim, C. *Chem. Rev.* **2003**, *103*, 773–787.

(4) Torrance, J. W.; MacArthur, M. W.; Thornton, J. M. *Proteins Struct. Funct. Bioinformatics* **2008**, *71*, 813–830.

(5) Schumacher, M.; Rivard, A.; Bachinger, H.; Adelman, J. *Nature* **2001**, *410*, 1120–1124.

(6) Sakharov, D.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921–4929.

(7) Bloomfield, V. *Biopolymers* **1991**, *31*, 1471–1481.

(8) Cate, J.; Gooding, A.; Podell, E.; Zhou, K.; Golden, B.; Kundrot, C.; Cech, T.; Doudna, J. *Science* **1996**, *273*, 1678–1685.

(9) Fang, X.; Pan, T.; Sosnick, T. *Nat. Struct. Biol.* **1999**, *6*, 1091–1095.

(10) Draper, D.; Grilley, D.; Soto, A. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 221–243.

(11) Woodson, S. *Curr. Opin. Chem. Biol.* **2005**, *9*, 104–109.

(12) Qu, X.; Smith, G. J.; Lee, K. T.; Sosnick, T. R.; Pan, T.; Scherer, N. F. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6602–6607.

(13) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.

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

(15) Jensen, K. P.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 1499–1509.

(16) Joung, I. S.; Cheatham, T. E. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(17) Joung, I. S.; Cheatham, T. E. *J. Phys. Chem. B* **2009**, *113*, 13279–13290.

(18) Carlsson, J.; Aqvist, J. *J. Phys. Chem. B* **2009**, *113*, 10255–10260.

(19) Roux, B. *Chem. Phys. Lett.* **1993**, *212*, 231–240.

(20) Allen, T. W.; Andersen, O. S.; Roux, B. *Biophys. J.* **2006**, *90*, 3447–3468.

(21) Bucher, D.; Guidoni, L.; Maurer, P.; Rothlisberger, U. *J. Chem. Theory Comput.* **2009**, *5*, 2173–2179.

(22) Yu, H. B.; Roux, B. *Biophys. J.* **2009**, *97*, L15–L17.

(23) Rick, S. W.; Stuart, S. J. *Rev. Comput. Chem.* **2002**, 89–146.

(24) Yu, H. B.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69–85.

(25) Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045.

(26) Lopes, P. E. M.; Roux, B.; MacKerell, A. D., Jr. *Theor. Chem. Acc.* **2009**, *124*, 11–28.

(27) Dang, L. X.; Rice, J. E.; Caldwell, J.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 2481–2486.

(28) Grossfield, A.; Ren, P.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.

(29) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2006**, *110*, 3308–3322.

(30) Jiao, D.; King, C.; Grossfield, A.; Darden, T.; Ren, P. *J. Phys. Chem. B* **2006**, *110*, 18553–18559.

(31) Drude, P. *Lehrbuch der Optik*; S. Hirzel: Leipzig, 1900.

(32) Sangster, M. J. L.; Dixon, M. *Adv. Phys.* **1976**, *25*, 247–342.

(33) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025–3039.

(34) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

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

(36) Dick, B. G., Jr.; Overhauser, A. W. *Phys. Rev.* **1958**, *112*, 90.

(37) Cochran, W. *Crit. Rev. Solid State Mater. Sci.* **1971**, *2*, 1–44.

(38) van Maaren, P. J.; van der Spoel, D. *J. Phys. Chem. B* **2001**, *105*, 2618–2626.

(39) Kunz, A. P.; van Gunsteren, W. F. *J Phys Chem A* **2009**, *113*, 11570–11579.

(40) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theor. Comput.* **2007**, *3*, 1120–1133.

(41) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2007**, *111*, 2873–2885.

(42) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theor. Comput.* **2007**, *3*, 1927–1946.

(43) Harder, E.; Anisimov, V. M.; Whitfield, T. W.; MacKerell, A. D., Jr.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509–3521.

(44) Lopes, P. E. M.; Lamoureux, G.; Mackerell, A. D., Jr. *J. Comput. Chem.* **2009**, *30*, 1821–1838.

(45) Harder, E.; MacKerell, A. D., Jr.; Roux, B. *J. Am. Chem. Soc.* **2009**, *131*, 2760–2761.

(46) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 1987.

(47) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990–2001.

(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) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341–350.

(50) van der Hoef, M. A.; Madden, P. A. *Mol. Phys.* **1998**, *94*, 417–433.

(51) Whitfield, T. W.; Martyna, G. J. *Chem. Phys. Lett.* **2006**, *424*, 409–413.

(52) Piquemal, J. P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 054511.

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

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

(55) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

(56) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(57) Martyna, G.; Tobias, D.; Klein, M. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

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

(59) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117–1157.

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

(61) Asthagiri, D.; Pratt, L. R.; Ashbaugh, H. S. *J. Chem. Phys.* **2003**, *119*, 2702–2708.

(62) Harder, E.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 234706.

(63) Kastenholz, M.; Hunenberger, P. *J. Chem. Phys.* **2006**, *124*, 124106.

(64) Kastenholz, M. A.; Huenenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 224501.

(65) Mahan, G. *Phys. Rev. A* **1980**, *22*, 1780–1785.

(66) Hattig, C.; Hess, B. A. *J. Chem. Phys.* **1998**, *108*, 3863–3870.

(67) Coker, H. *J. Phys. Chem.* **1976**, *80*, 2078–2084.

(68) Coker, H. *J. Phys. Chem.* **1976**, *80*, 2084–2091.

(69) Pyper, N. C.; Pike, C. G.; Edwards, P. P. *Mol. Phys.* **1992**, *76*, 353–372.

(70) Jemmer, P.; Fowler, P. W.; Wilson, M.; Madden, P. A. *J. Phys. Chem. A* **1998**, *102*, 8377–8385.

(71) Džidič, I.; Kebarle, P. *J. Phys. Chem.* **1970**, *74*, 1466–1474.

(72) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R., Jr. *J. Phys. Chem. A* **1998**, *102*, 7787–7794.

(73) Feller, D.; Glendening, E. D.; Kendall, R. A.; Peterson, K. A. *J. Chem. Phys.* **1994**, *100*, 4981–4997.

(74) Glendening, E. D.; Feller, D. *J. Phys. Chem.* **1995**, *99*, 3060–3067.

(75) Borodin, O.; Bell, R. L.; Li, Y.; Bedrov, D.; Smith, G. D. *Chem. Phys. Lett.* **2001**, *336*, 292–302.

(76) Kim, J.; Lee, H. M.; Suh, S. B.; Majumdar, D.; Kim, K. S. *J. Chem. Phys.* **2000**, *113*, 5259.

(77) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; 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.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(78) Randles, J. *Trans. Faraday Soc.* **1956**, *52*, 1573–1581.

(79) Noyes, R. M. *J. Am. Chem. Soc.* **1962**, *84*, 513–522.

(80) Gomer, R.; Tryson, G. *J. Chem. Phys.* **1977**, *66*, 4413–4424.

(81) Klots, C. E. *J. Phys. Chem.* **1981**, *85*, 3585–3588.

(82) Marcus, Y. *J. Chem. Soc., Faraday Trans.* **1991**, *87*, 2995–2999.

(83) Schmid, R.; Miah, A. M.; Sapunov, V. N. *Phys. Chem. Chem. Phys.* **2000**, *2*, 97–102.

(84) Whitfield, T. W.; Varma, S.; Harder, E.; Lamoureux, G.; Rempe, S. B.; Roux, B. *J. Chem. Theory Comput.* **2008**, *3*, 2068.

(85) Wagman, D. D.; Evans, W. H.; Parker, V. B.; Schumm, R. H.; Halow, I.; Bailey, S. M.; Churney, K. L.; Nuttall, R. L. *J. Phys. Chem. Ref. Data* **1982**, *11*, Supplement No. 2.

(86) Peschke, M.; Blades, A. T.; Kebarle, P. *J. Phys. Chem. A* **1998**, *102*, 9978–9985.

(87) Peschke, M.; Blades, A. T.; Kebarle, P. *J. Am. Chem. Soc.* **2000**, *122*, 10440–10449.

(88) Ohtaki, H.; Radnai, T. *Chem. Rev.* **1993**, *93*, 1157–1204.

(89) Spangberg, D.; Rey, R.; Hynes, J. T.; Hermansson, K. *J. Phys. Chem. B* **2003**, *107*, 4470–4477.

(90) Varma, S.; Rempe, S. B. *Biophys. Chem.* **2006**, *124*, 192–199.

(91) Lyubartsev, A. P.; Laasonen, K.; Laaksonen, A. *J. Chem. Phys.* **2001**, *114*, 3120–3126.

(92) Todorova, T.; Hunenberger, P. H.; Hutter, J. *J. Chem. Theory Comput.* **2008**, *4*, 779–789.

(93) Caminiti, R.; Licheri, G.; Piccaluga, G.; Pinna, G. *Chem. Phys. Lett.* **1977**, *47*, 275–278.

(94) Martinez, J. M.; Pappalardo, R. R.; Marcos, E. S. *J. Am. Chem. Soc.* **1999**, *121*, 3175–3184.

(95) *CRC Handbook of Chemistry and Physics*, 87th ed.; Lide, D. R., Ed.; Taylor and Francis: Boca Raton, FL, 2007.

(96) Roux, B. *Biophys. J.* **1996**, *71*, 3177–3185.

(97) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587–1597.

(98) Luo, Y.; Roux, B. *J. Phys. Chem. Lett.* **2010**, *1*, 183–189.

CT900576A

JCTC Journal of Chemical Theory and Computation

Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models

Gregory R. Bowman,[†] Daniel L. Ensign,[‡] and Vijay S. Pande*[,†,‡]

Biophysics Program and Department of Chemistry, Stanford University, Stanford, California 94305

Received November 22, 2009

Abstract: Computer simulations can complement experiments by providing insight into molecular kinetics with atomic resolution. Unfortunately, even the most powerful supercomputers can only simulate small systems for short time scales, leaving modeling of most biologically relevant systems and time scales intractable. In this work, however, we show that molecular simulations driven by adaptive sampling of networks called Markov State Models (MSMs) can yield tremendous time and resource savings, allowing previously intractable calculations to be performed on a routine basis on existing hardware. We also introduce a distance metric (based on the relative entropy) for comparing MSMs. We primarily employ this metric to judge the convergence of various sampling schemes but it could also be employed to assess the effects of perturbations to a system (e.g., determining how changing the temperature or making a mutation changes a system's dynamics).

1. Introduction

Molecular dynamics simulations are a powerful means of understanding both the thermodynamics and kinetics of molecular processes like protein folding and conformational changes. Unfortunately, such processes are highly sensitive to the underlying chemical details. For example, point mutations in the amino acid sequence of a protein may have significant effects on its kinetics,[1] and a small number of point mutations can even drastically change the native structure.[2] Thus, atomistic simulations are required to make quantitative connections with experiments.[3,4]

Advances in computing have made it possible to rapidly generate huge data sets even at this level of chemical detail;[5,6] however, these data sets are still insufficient. A typical computer can only simulate ∼5 ns/day of protein folding and would thus take over 500 years to simulate one millisecond, an average folding time typical of proteins. Whether one is interested in dynamics or merely equilibrium probabilities, a kinetic perspective on this problem that explicitly considers the rate of equilibration reveals that metastability, or the presence of long-lived states that act as "traps", is a common source of inefficiency.

One approach to dealing with this issue is to make tremendous investments in specialized software and hardware for generating long simulations.[7] While theoretically sound,[8] this serial approach often only results in simulations that are long *relative* to standard trajectories. However, a *truly long* simulation must be orders of magnitude longer than the slowest relaxation time so that the probabilities of all states and pathways can be estimated accurately. Even if such a simulation were possible, the task of analyzing the data would still remain.[7,9] Moreover, serial approaches are inherently inefficient, both due to parallelization overhead and, more importantly, the fact that they waste hundreds of years of computing time waiting for rare events.

A statistical approach provides a fundamentally different perspective on model construction. Rather than attempting to generate one realization of an entire process, one instead aims to generate an ensemble of events in parallel. For example, a number of methods have been developed for exploiting statistical mechanics to simulate protein folding more efficiently.[10−13] Most of these approaches rely on the fact that, in two-state protein folding, the waiting time for observing a transition is exponentially distributed but the actual transition times are quite rapid.[14] Thus, proteins often

* Corresponding author e-mail: pande@stanford.edu.
[†] Biophysics Program.
[‡] Department of Chemistry.

10.1021/ct900620b
Published on Web 02/17/2010

fold much faster or slower than the average folding time. Such approaches are amenable to commodity hardware and take far less wall-clock time than a serial approach with an equivalent amount of sampling, particularly when combined with grid computing.[5] Unfortunately, these methods are generally only applicable to two-state systems and may require simulations of an unknown minimum length.[15] Some multistate generalizations exist[16] but quickly become computationally intractable.

Markov state models (MSMs) extend this work by allowing for a tractable, multistate scheme that allows efficient modeling of any system exhibiting metastability.[17] A MSM is a network with nodes corresponding to metastable states and edges describing the rates of transitioning between pairs of states, akin to a map with cities connected by roads labeled with speed limits. Rather than attempting to generate one realization of an entire process, one can exploit the decomposition of conformational space into multiple metastable states to gather statistics on each step of the process independently, allowing a problem to be broken up into more manageable and trivially parallelizable pieces.

Mathematically, MSMs are represented as transition probability matrices, with the entry in row i and column j giving the probability of transitioning from state i to state j within a time interval called the lag time of the model. Building MSMs is a challenging task, but significant progress has been made over the past few years,[18−21] leading to freely available software for automatically constructing these models.[18] While MSMs could be used to analyze truly long simulations, their ultimate value lies in their ability to facilitate efficient model construction by allowing precise, parallel determination of the transition rates between states by running many short simulations from each of them.

Adaptive sampling algorithms for MSM construction take this statistical approach a step further.[22−24] In adaptive sampling, one first obtains an initial model of the entire process of interest by any means possible. One then iteratively calculates the contribution of each step of the process to uncertainties in some observable of interest via Bayesian statistics and runs numerous parallel simulations of the steps that can lead to the greatest increases in precision until the desired level of statistical certainty is achieved. Such an approach was recently shown to lead to dramatic reductions in the statistical uncertainty in the observable of interest relative to other refinement schemes.[22]

However, a number of important questions remain to be answered. First, does adaptive sampling improve the global model quality or just local components that are important for the observable of interest? Exactly how much more efficient is adaptive sampling? And finally, is adaptive sampling capable of discovering previously unknown components of a model, or is it only able to refine the initial model it is given?

In this work, we address these questions using a MSM for the villin headpiece (HP-35 NleNle) that was recently constructed from atomistic simulations with explicit solvent.[19] We then move on to simple models, where the role of the network is clear, to gain an intuition for our results and test whether such methods could be more broadly applicable to a wide class of different types of systems. These analyses rely on a new distance metric for MSMs developed in section 2.2, which should prove generally useful for evaluating various sampling schemes and even assessing the effects of perturbations to a system (like changes in temperature or even mutations).

2. Theoretical Underpinnings

2.1. Adaptive Sampling. In adaptive sampling approaches to MSM construction, simulations are run iteratively to minimize uncertainties in some property of a model.[22−24] In this work, adaptive sampling is performed as follows:

(1) Perform N simulations of L steps starting from a particular starting state(s).

(2) Build a MSM only including those states identified so far.

(3) Calculate the contribution of each state to uncertainty in the slowest kinetic rate following ref 22.

(4) Start N new simulations of L steps distributed among the states in proportion to their contribution to uncertainty in the slowest rate.

(5) Repeat steps 2−4 for some number of iterations.

All the MSMs in this work were constructed and analyzed with the MSMBuilder package (which is freely available at https://simtk.org/home/msmbuilder/)[18] modified such that transition count matrices were not symmetrized by counting the transitions that would have been observed if one watched each simulation backward.

We note that, in the past, simulations in each round of adaptive sampling were all started from the same initial state (the one contributing most to uncertainty in the quantity of interest).[22] The intuition behind our alteration was that, as the number of simulations (N) becomes large, starting all the simulations from one state would be excessive as fewer would be sufficient to drastically reduce the uncertainty. Instead, it would be preferable to allocate some of these excess simulations to reduce uncertainties in other states' transition probabilities. Indeed, we have found that our modified procedure yields better results for sufficiently large N on reasonably complex networks and gives equivalent results for simple networks and small N.

To demonstrate the utility of this algorithm, we carried out adaptive sampling with synthetic trajectories generated from transition count matrices. To generate synthetic simulations from a transition count matrix, we first normalize each row to obtain a transition probability matrix. At each time step (or each lag time), the next state is chosen according to the distribution of transition probabilities for the current state. The prior described below is not used for these calculations, so the matrices used to generate trajectories tend to be sparse.

2.2. Quantifying the Similarity between MSMs. In order to monitor the convergence of any sampling scheme, it is important to first develop a similarity metric that is capable of measuring the global quality of a test model relative to some reference model. Such a metric would also have broad usefulness, as there are several reasons for comparing MSMs quantitatively. For example, this metric could be used to compare MSMs generated by two different simulation

methods, allowing one to directly compare the resulting dynamics. Alternatively, one could compare MSMs generated by two somewhat different, but related systems, such as comparing the simulations of the dynamics of two point mutants of a given protein.

We have developed such a distance metric for MSMs that is based on the relative entropy, which is a common measure of the distance between two probability distributions in information theory[25] with important physical implications.[26] The relative entropy between two normalized distributions P and Q, over a common set of outcomes, is

$$D(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$$

where P_i is the probability of outcome i, P is a reference distribution, and Q is some test distribution.

A MSM consists of one normalized distribution per state, which gives the probability of transitioning to each other state within one lag time. We define the relative entropy between a reference and test MSM, with transition matrices P and Q respectively, as

$$D(P||Q) = \sum_{i,j}^{N} P_i P_{ij} \log \frac{P_{ij}}{Q_{ij}} \quad (1)$$

where P_i is the equilibrium probability of state i, P_{ij} is the probability of transitioning from state i to state j during one lag time, and N is the number of states. Intuitively, our relative entropy metric is the sum of the relative entropies between the transition probability distributions for each state weighted by their stationary probabilities.

One may derive our relative entropy metric for MSMs more formally by considering that the entropy (H) of a sample path of a stochastic process, normalized by its length, is also called the entropy rate. An important theorem in information theory is the following:

Theorem. For an ergodic stochastic process, X_1, ..., X_n

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n) = \lim_{n\to\infty} H(X_n|X_1, ..., X_{n-1})$$

For a Markov Chain, the right-hand side takes a very simple form, because the conditional entropy only depends on the previous step, which converges to the stationary distribution.

In the following, we prove a similar statement for the relative entropy between the paths of two Markov chains as n goes to infinity. For two Markov chains p and q with state space Ω, we would like to compute:

$$\lim_{n\to\infty} \frac{1}{n} D(p(X_1, ..., X_n)||q(X_1, ..., X_n))$$

For simplicity, let us define lowercase $x_n = \{X_1, ..., X_n\}$. Then, by the chain rule for the relative entropy, we get

$$\lim_{n\to\infty} \frac{1}{n}[D(p(x_{n-1})||q(x_{n-1})) + D(p(X_n|x_{n-1})||q(X_n|x_{n-1}))] \quad (2)$$

Equation 2.65 in Cover and Thomas[27] defines the conditional relative entropy above as the expectation of the relative entropy between the conditional distributions of X_n given x_{n-1}, with respect to the distribution of x_{n-1}. This means that

$$\begin{aligned} D(p(X_n|x_{n-1})||q(X_n|x_{n-1})) &= \sum_{y\in\Omega^{n-1}} p(x_{n-1} = y) D(p(X_n|y)||q(X_n|y)) \\ &= \sum_{Y\in\Omega} p(X_{n-1} = Y) D(p(X_n|Y)||q(X_n|Y)) \end{aligned}$$

where we have grouped terms with the same final state in the "history" y, which have the same relative entropy factor, and summed their probabilities to obtain the marginal probability over X_{n-1}.

Repeating the step that led to eq 2 many times yields

$$\lim_{n\to\infty} \frac{1}{n}[\sum_{m=2}^{n} D(p(X_m|x_{m-1})||q(X_m|x_{m-1}))] + D(p(X_1)||q(X_1))$$

If the initial state is deterministic, the last term is just zero. As for the first term, as n goes to infinity, the distribution of X_{m-1} goes to the stationary distribution of p, which we call μ. Then, using the equation for the conditional entropy,

$$\lim_{n\to\infty} D\left(p(X_n\middle|x_{n-1})\middle||q(X_n\middle|x_{n-1})\right) = \sum_{Z\in\Omega} \mu(Z) \sum_{Y\in\Omega} p(Y|Z) \log\left[\frac{p(Y|Z)}{q(Y|Z)}\right]$$

Since the terms in the series converge to a limit, their Cesaro means converge to the same limit, so

$$\lim_{n\to\infty} \frac{1}{n} D(p(X_1, ..., X_n)||q(X_1, ..., X_n)) = \sum_{Z\in\Omega} \mu(Z) \sum_{Y\in\Omega} p(Y|Z) \log\left[\frac{p(Y|Z)}{q(Y|Z)}\right]$$

The terms $p(Y|Z)$ and $q(Y|Z)$ are just the elements of the transition matrices of p and q, respectively; so this is equivalent to eq 1.

2.3. Prior for Relative Entropy and Adaptive Sampling. There is always some probability of transitioning between every pair of states, though these probabilities may be low enough that no actual transitions are observed. To account for this, as well as to reflect our lack of prior knowledge about the transition probabilities, we add a pseudocount of $1/N$ to every element of the transition count matrix, where N is the number of states, before normalizing each row to find the transition probability matrix, as in refs 22 and 28. The intuition behind this choice is that for a state to exist we must observe at least one count in that state, but before observing any real data the probability of this count leading to any other state is equal. From a Bayesian perspective, these pseudocounts equate to a uniform prior. These pseudocounts also prevent the relative entropy metric from becoming infinite whenever a zero is encountered in a MSM's transition probability matrix. It is often the case that certain transitions are not observed, so this correction is of great practical importance.

2.4. Villin Simulations and MSM. The simulation details for the original ~450 villin simulations are described in detail

in ref 29. In short, ~450 constant temperature molecular dynamics simulations with explicit solvent and up to 2 μs in length were run from nine initial configurations drawn from high temperature unfolding simulations at 373 K. Ref 19 describes the construction of a 10 000 microstate MSM that faithfully reproduces the raw simulation data. For the purposes of this work, we lumped these 10 000 microstates into 500 macrostates exhibiting metastability and having an equivalent Markov time (15 ns). This lumping was done with the MSMBuilder package.[18] The macrostates containing the nine initial configurations used during the real simulations were used as the starting points for adaptive sampling. Simulations of just 30 ns were used for adaptive sampling.

2.5. Simple Models. The transition count matrices for simple models S and P (C_S and C_P respectively) are

$$C_S = \begin{bmatrix} 6000 & 3 & 0 & 0 & 0 & 0 \\ 3 & 1000 & 3 & 0 & 0 & 0 \\ 0 & 3 & 1000 & 3 & 0 & 0 \\ 0 & 0 & 3 & 1000 & 3 & 0 \\ 0 & 0 & 0 & 3 & 1000 & 3 \\ 0 & 0 & 0 & 0 & 3 & 90\,000 \end{bmatrix}$$

and

$$C_P = \begin{bmatrix} 6000 & 2 & 2 & 0 & 0 & 0 \\ 2 & 1000 & 0 & 2 & 2 & 0 \\ 2 & 0 & 1000 & 2 & 2 & 0 \\ 0 & 2 & 2 & 1000 & 0 & 2 \\ 0 & 2 & 2 & 0 & 1000 & 2 \\ 0 & 0 & 0 & 2 & 2 & 90\,000 \end{bmatrix}$$

where the entry in row i and column j gives the number of transitions observed from state i to state j.

Mean first passage times were calculated following ref 28. The mean first passage times for S and P are ~13 000 and ~5000 steps, respectively. Other equilibrium properties can be obtained by normalizing each row to obtain a transition probability matrix and then solving for the eigenvalues and eigenvectors of this matrix. For example, normalizing the first eigenvector (e.g., the one corresponding to an eigenvalue of 1) gives the equilibrium probabilities of each state. Subsequent eigenvalue/eigenvector pairs give kinetic rates and the states involved in these transitions, respectively.[17] Once again, the MSMBuilder package[18] was used for analysis of these models.

Plots of the average relative entropy as a function of simulation number and length were generated by running 600 simulations of 5000 steps for each model. Average relative entropies over 10 random samples of N trajectories from this pool were then calculated and plotted. Similar plots for our adaptive sampling scheme were also generated by averaging over 10 independent runs.

3. Results and Discussion

3.1. Application to Villin MSM. With these tools in place, we are now in a position to assess the efficacy of adaptive sampling using a previously calculated MSM for the villin headpiece[19] as a model system. In particular, we would like to assess two types of efficiency. First, given our desire to push the envelope of what is possible in a reasonable amount of time, can adaptive sampling reduce the wall-clock time necessary to achieve a given model quality? Second, given our desire to mitigate negative impacts on the environment, can adaptive sampling reduce the amount of resources (in this case computer time) necessary to achieve a given model quality?

To address these questions, we have performed adaptive sampling with a variable number of simulations per iteration generated from our villin MSM. We then assume each simulation progresses at a rate of 5 ns/day, a typical value for modern personal computers, and compare the convergence of our adaptive simulations to the gold-standard model from ref 19 (that was validated by comparison to both the raw simulation data and experiments) with the convergence of a single long reference simulation to the same gold standard. Convergence to the gold-standard model is measured with our relative entropy metric for MSMs (described in section 2.2).

Figure 1A shows that the wall-clock time efficiency of adaptive sampling scales linearly up to 5000 simulations per iteration. That is, adaptive sampling with N simulations per iteration can reduce the wall-clock time necessary to achieve a given model quality by a factor of N for N as high as 5000. Using more simulations will help but will only reduce the wall-clock time by a factor of αN, where $\alpha < 1$. The crucial result, however, is that one can reduce a calculation that would take decades to run with traditional methods to a calculation that can be run in a matter of days with adaptive sampling.

Adaptive sampling can also greatly reduce the resource requirements for achieving a given model quality. For example, Figure 1B shows the computer time necessary to achieve a given model quality for one long simulation and adaptive sampling with a varying number of simulations per iteration. This figure shows that adaptive sampling requires about half as much computer time to achieve the same model quality as one long simulation. Once again, the relative efficiency of adaptive sampling begins to fall off beyond some optimal number of simulations per iteration.

3.2. Application to Simple Models. To gain intuition for the applicability of adaptive sampling to other systems, we have also applied it to two classic network topologies, shown in Figure 2A and defined more thoroughly in section 2.5. These models are representative of problems with metastability; their equilibrium properties can be derived analytically and used as an unambiguous reference, and truly long simulations are feasible.

Both models have states with approximately the same equilibrium and transition probabilities, such that differences between their behaviors can be attributed to differences between their topologies. More specifically, states 1−6 have equilibrium populations of 6%, 1%, 1%, 1%, 1%, and 90%, respectively. Drawing an analogy to protein folding, state 1 is the unfolded state, state 6 is the folded state, and the remaining states are intermediates. Thus, S has a single folding pathway, and P has parallel folding pathways.

The reduced connectivity in S results in longer time scale transitions relative to P. In fact, the mean first passage time

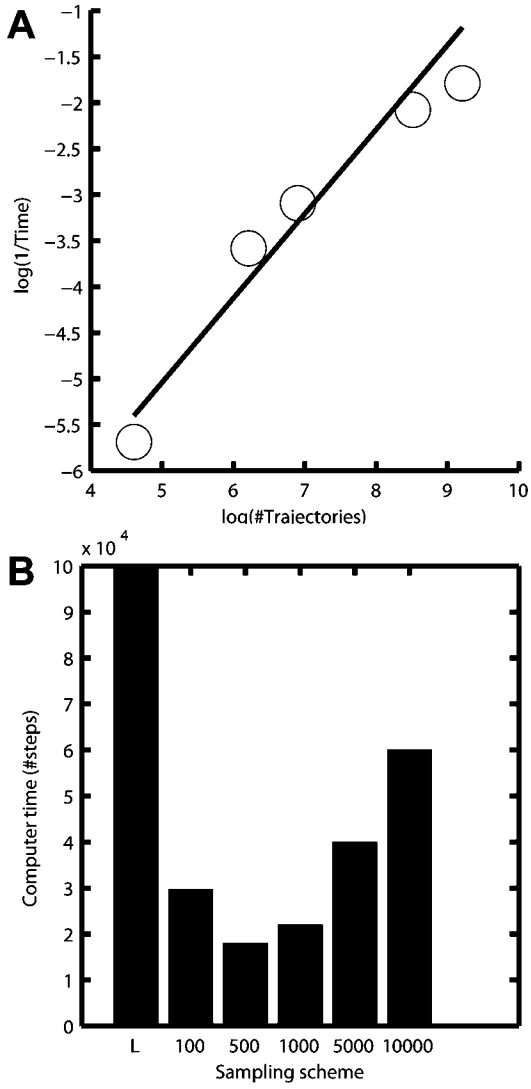


Figure 1. Scaling for adaptive sampling of villin as the number of parallel simulations (*N*) used during each round is varied. (A) Wall-clock time scaling as *N* is varied. The black line is a best fit to the linear portion of the data (circles), which extends up to 5000 simulations per iteration. (B) Computer time required to achieve a given model quality (relative entropy) for various sampling schemes. *L* refers to one long trajectory, and the numbers refer to the number of parallel simulations used in each iteration of adaptive sampling. All results come from averaging over 10 independent runs. Each step equates to 15 ns.

(MFPT) between states 1 and 6 is about three times longer in S than in P, making S considerably harder to sample. In addition, such linear models are often cited as a case where the holistic, long-trajectory approach is absolutely necessary; nevertheless, adaptive sampling is able to learn the network more efficiently than traditional approaches, as shown in Figure 2B. This figure shows how close various schemes can approach the true model for S given a set amount of wall-clock time and starting from state 1 to mimic the practice of starting protein folding simulations from an arbitrary conformation in the unfolded state.

To provide some intuition for our distance metric, Figure 3 shows the evolution of the relative entropy and the estimated free energy of each state in S during adaptive

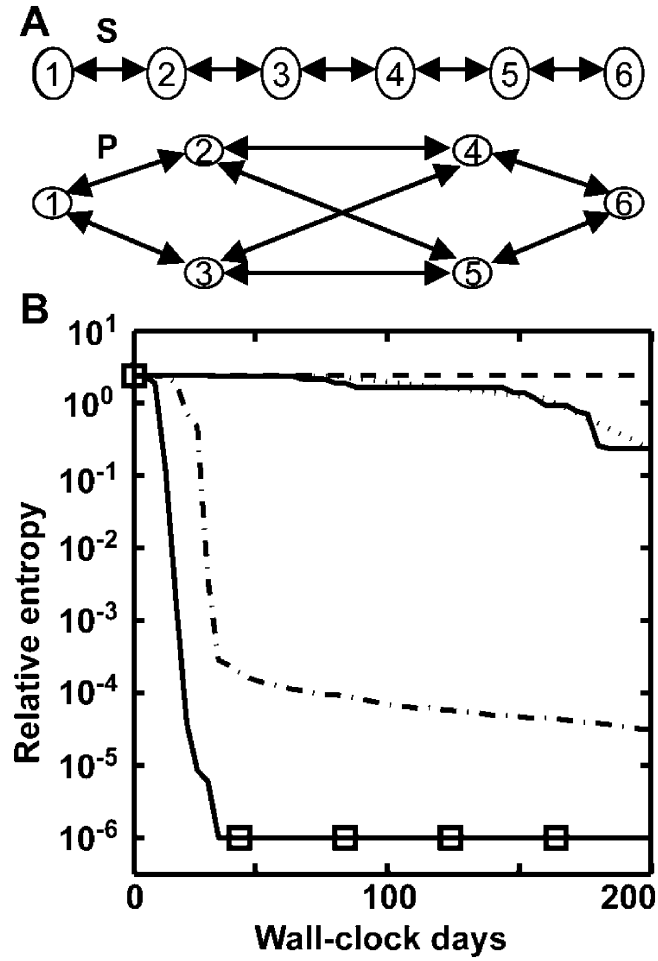


Figure 2. (A) The two models, S and P. (B) Distance from the true model (measured via the relative entropy) as a function of wall-clock time for adaptive sampling versus one long simulation of S (assuming 5 steps/day to mimic 5 ns/day in protein folding simulations). The lines are one long simulation (dashed line) and adaptive sampling with 10 simulations of 20 steps (solid line), 10 simulations of 200 steps (dotted line), 100 simulations of 20 steps (dash-dot line), and 1000 simulations of 20 steps (black squares) per iteration.

sampling. Adaptive sampling was carried out by running 10 simulations from state 1 and then repeatedly building a MSM and starting 10 new simulations from the state contributing most to uncertainty in the slowest process. Small jumps in the relative entropy are found each time a state with a low population is discovered (or, equivalently, when a new path is discovered for this model), and a very large jump is evident when the most populated state, state 6, is discovered. Slow decay occurs between these jumps. Thus, our metric is most sensitive to state and path discovery but still captures improvements in estimates of the transition probabilities along known paths. Such behavior is desirable as models that miss important states or paths should be penalized more than ones with imperfect transition probabilities.

Figure 4 shows a more thorough comparison of adaptive sampling and reference simulations with an equal amount of sampling for various numbers and lengths of simulations. Evaluation of the reference simulations for both S and P demonstrates that achieving a reasonable model quality by naively starting simulations from state 1 requires simulations

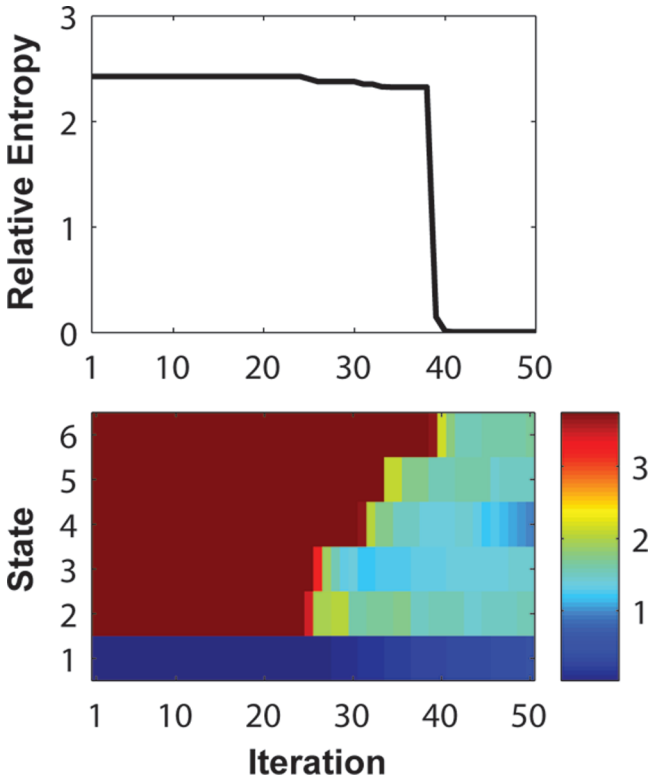


Figure 3. Relative entropy (top) and free energy of each state in kcal/mol (bottom) as a function of the adaptive sampling iteration on model S.

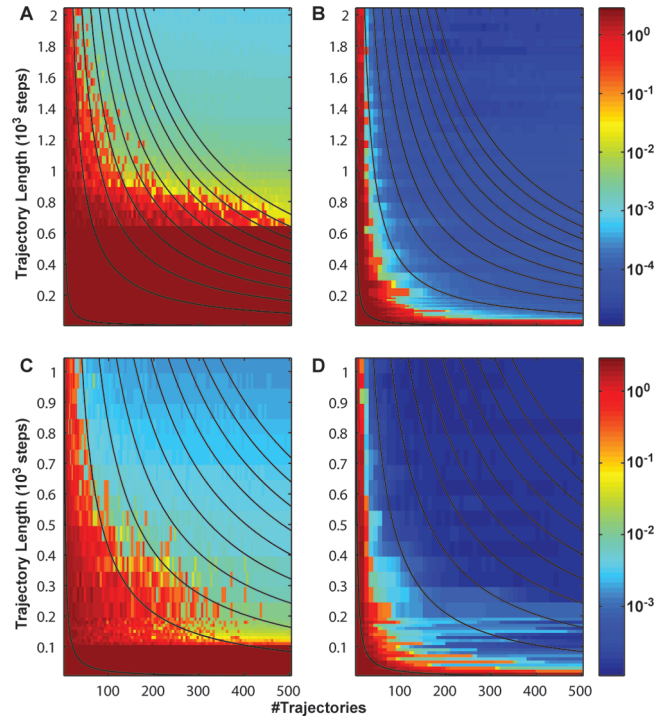


Figure 4. Distance from the true model (measured via the relative entropy) as a function of the number and length of simulations averaged over 10 independent samples. (A) Reference distribution for S, (B) adaptive sampling of S, (C) reference distribution for P, and (D) adaptive sampling of P. All simulations for the reference distributions started from state 1. The first 10 simulations for adaptive sampling started from state 1, and subsequent batches of simulations started from the state contributing most to uncertainty in the slowest process. Black lines are contours of equal amounts of data.

of some minimal length, though this minimal length is shorter for P than S in terms of the absolute number of steps. Moreover, adaptive sampling is able to gain valuable information from much shorter and fewer simulations regardless of the topology of the network, that is, whether there is a single folding pathway or multiple pathways. This figure also shows that adaptive sampling generally benefits from using more parallel simulations but not longer ones. An important point is that each data point in Figure 4B and D depends on the data points to its left. For example, to fill in the row corresponding to simulations of length 100, 10 independent adaptive sampling runs of 50 iterations were performed. The first round of each adaptive sampling run was used to compute average relative entropies for 1−10 simulations, the first and second round of each run (which depends on the first round) for 11−20 simulations, and so forth. As a result, there is some horizontal streakiness in these figures. We also note that adaptive sampling results in smaller uncertainties in the relative entropies shown in Figure 4 (see Figures S1 and S2, Supporting Information).

Finally, we find that the scaling of adaptive sampling of our simple networks is similar to that found for villin, as shown in Figure 5. One noteworthy difference is that our simple models saturate (i.e., fall short of linear scaling as additional parallel simulations are run) earlier than villin. Comparison of the two simple models also shows that S saturates before P. For S, adaptive sampling scales linearly up to 150 parallel simulations. For P, adaptive sampling scales linearly up to 500 simulations. The improved scaling for P is the result of the increased complexity of the network topology of P compared to S. Each node in P has more connections to learn, and the algorithm benefits from doing this in parallel. Indeed, the complexity of our villin model is much greater than either of these simple networks, and as discussed previously, villin scales linearly up to 5000 simulations per iteration. Thus, we expect that we can achieve linear scaling well beyond 5000 simulations per iteration for systems that are more complex than the villin MSM that we sampled from.

3.3. Applicability. The adaptive sampling algorithm employed here was developed for application to MSMs with metastable states. That is, it assumes that every state has a self-transition probability greater than 0.5 such that a simulation in one state is more likely to stay there than to transition to a new state. This property helps to ensure a separation of time scales (fast intrastate transitions, slow interstate transitions) and, therefore, that the model is Markovian because a simulation can lose memory of its previous state before transitioning to a new one. Thus, the procedure for *ab initio* adaptive sampling is (1) run some initial simulations, (2) cluster all the simulation data into microstates, (3) lump these microstates into metastable macrostates, (4) calculate the contribution of each macrostate to uncertainties in the slowest rate (or some other observable), (5) start new simulations from each state in proportion to its contribution to the overall uncertainty, and (6) repeat steps 2−5 until the desired level of statistical certainty is achieved.

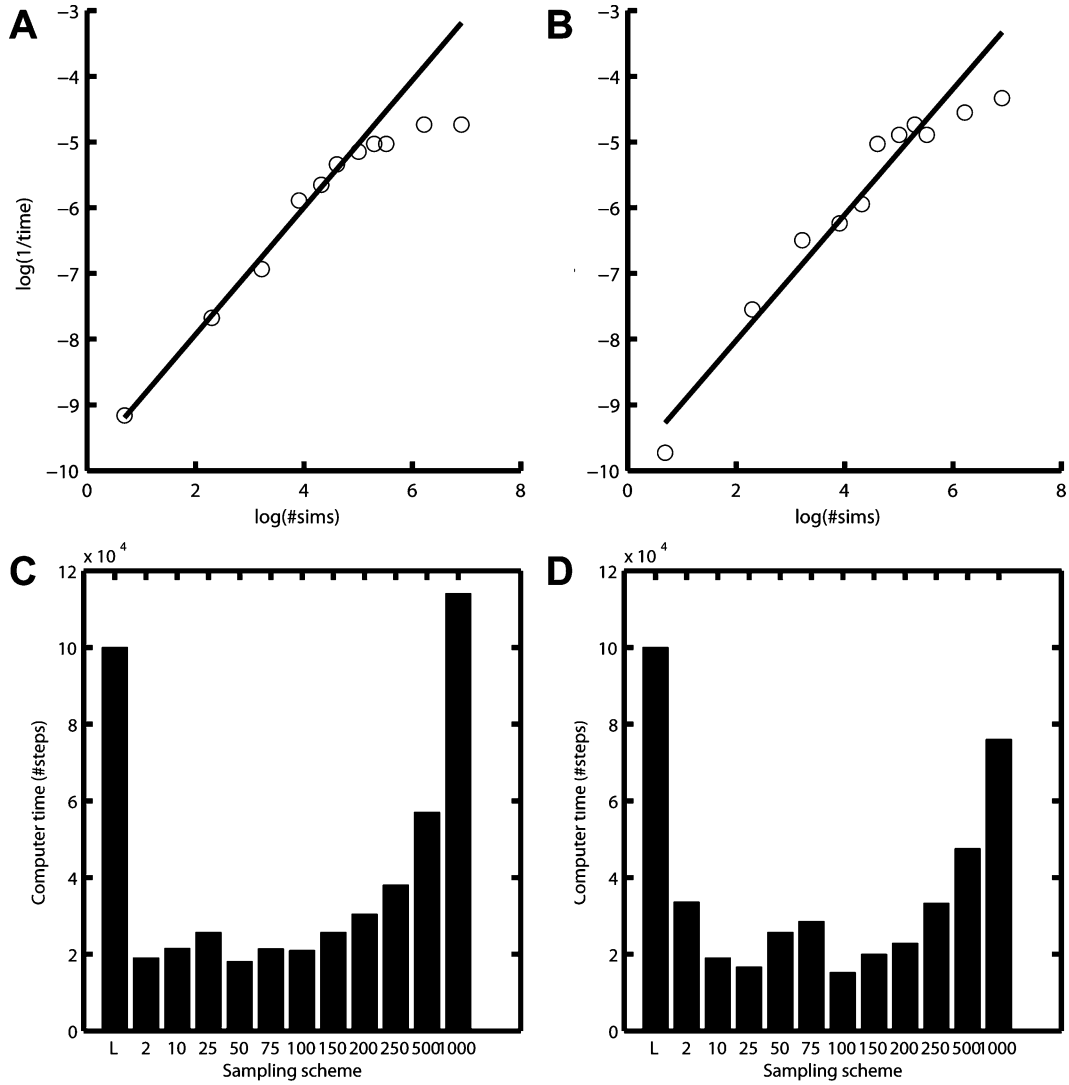


Figure 5. Scaling for adaptive sampling of our simple models as the number of parallel simulations (*N*) used during each round is varied. (A and B) Wall-clock time scaling as *N* is varied for simple models S and P, respectively. The black line is a best fit to the linear portion of the data (circles). (C and D) Computer time required to achieve a given model quality (relative entropy) for various sampling schemes applied to S and P, respectively. *L* refers to one long trajectory, and the numbers refer to the number of parallel simulations used in each iteration of adaptive sampling. All results come from averaging over 10 independent runs.

In the future, it will be interesting to explore whether this adaptive sampling algorithm is equally applicable to more fine grained divisions of conformational space (e.g., at the microstate level) as the lumping stage would no longer be necessary. In addition, recent work has shown that more fine grained MSMs are better for obtaining quantitative predictions of experimental observables,[19,30,31] so it could be advantageous to do refinement at this level.

The relative entropy metric assumes that the two models being compared have the same state space. Comparing two simulation data sets therefore requires the following steps: (1) define a state space common to both data sets (i.e., by using both data sets for clustering to define microstates and, optionally, lumping to define macrostates), (2) compute transition probability matrices for each data set independently, and (3) compute the relative entropy between these matrices.

4. Conclusions

Together, our results with villin and fundamental model systems demonstrate the tremendous value of adaptive sampling. Since model quality has been assessed with a global metric and shows strong agreement between adaptive sampling results and the true model, we can conclude that adaptive sampling to minimize uncertainties in the slowest kinetic rate improves the global quality of a model. Moreover, adaptive sampling is significantly more efficient than a single long simulation, both in terms of the wall-clock time and resources required to achieve a given model quality, up to some saturation point. In fact, adaptive sampling with *N* parallel simulations requires about a factor of 2 less computer-time and a factor of *N* less wall-clock time. Considering that *N* can easily be as large as 10 000 (or more),[5] this can be a truly dramatic advantage in wall-clock time, turning calculations normally requiring decades into

routine calculations on the time scale of days. Finally, since our simulations started from just a couple of states, we can conclude that adaptive sampling is capable of discovering new model components *given no prior knowledge of the system* and is thus useful for model construction in addition to model refinement.

The adaptive sampling method described here may be directly applied to learn models from simulations of metastable phenomena, leading to significant resource and time savings in fields like molecular and quantum mechanics, but is not limited to these applications. Given a means to prepare samples within a given state, it could be applied equally well to experimental techniques, such as single molecule FRET and force extension experiments. More broadly, minimizing uncertainties in a model is likely to prove valuable even when metastability is not present. Similar methods may also be useful for understanding other complex network dynamics, as in signaling pathways.

Acknowledgment. Thanks to Sergio Bacallado for help with the relative entropy metric. This work was funded by NIH R01-GM062868 and NIH U54 GM072970. G.R.B. was supported by the NSF GRFP.

Supporting Information Available: Figures S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 2369–2374.

(2) He, Y.; Yeh, D. C.; Alexander, P.; Bryan, P. N.; Orban, J. *Biochemistry* **2005**, *44*, 14055–14061.

(3) Rhee, Y. M.; Pande, V. S. *J. Chem. Phys.* **2006**, *323*, 66–77.

(4) Bradley, P.; Misura, K. M.; Baker, D. *Science* **2005**, *309*, 1868–1871.

(5) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903–1904.

(6) Das, R.; Qian, B.; Raman, S.; Vernon, R.; Thompson, J.; Bradley, P.; Khare, S.; Tyka, M. D.; Bhat, D.; Chivian, D.; Kim, D. E.; Sheffler, W. H.; Malmstrom, L.; Wollacott, A. M.; Wang, C.; Andre, I.; Baker, D. *Proteins* **2007**, *69*, 118–128.

(7) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(8) Geyer, C. J. *Stat. Sci.* **1992**, *7*, 473–511.

(9) King, R. D.; Rowland, J.; Oliver, S. G.; Young, M.; Aubrey, W.; Byrne, E.; Liakata, M.; Markham, M.; Pir, P.; Soldatova, L. N.; Sparkes, A.; Whelan, K. E.; Clare, A. *Science* **2009**, *324*, 85–89.

(10) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91–109.

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

(12) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880–10889.

(13) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983–4987.

(14) Chung, H. S.; Louis, J. M.; Eaton, W. A. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 11837–11844.

(15) Fersht, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 14122–14125.

(16) Rogal, J.; Bolhuis, P. G. *J. Chem. Phys.* **2008**, *129*, 224107.

(17) Schutte, C. Ph.D. Thesis, Freie Universitat, Berlin, 1999.

(18) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197–201.

(19) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.

(20) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.

(21) Noe, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

(22) Hinrichs, N. S.; Pande, V. S. *J. Chem. Phys.* **2007**, *126*, 244101.

(23) Roblitz, S. Ph.D. Thesis, Freie Universitat, Berlin, 2008.

(24) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19765–19769.

(25) MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, U. K., 2003.

(26) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.

(27) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*, 2nd ed.; Wiley-Inter Science: Hoboken, NJ, 2006.

(28) Singhal, N.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 204909.

(29) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806–816.

(30) Sarich, M.; Noe, F.; Schutte, C. *SIAM Multiscale Model. Simul.* In submission, **2010**.

(31) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011–19016.

CT900620B

JCTC Journal of Chemical Theory and Computation

Development of a Polarizable Force Field Using Multiple Fluctuating Charges per Atom

Dong-Xia Zhao,[†] Cui Liu,[†] Fang-Fang Wang,[†] Chun-Yang Yu,[†] Li-Dong Gong,[†] Shu-Bin Liu,[‡] and Zhong-Zhi Yang*,[†]

Chemistry and Chemical Engineering Faculty, Liaoning Normal University, Dalian, 116029, China and Research Computing Center, University of North Carolina, Chapel Hill, North Carolina 27599-3420

Received December 10, 2009

Abstract: A polarizable force field (PFF) using multiple fluctuating charges per atom, ABEEM$\sigma\pi$ PFF, is presented in this work. The fluctuating partial charges are obtained from the electronegativity equalization principle applied to the decomposition scheme of atom-bond regions into multiple charge sites: atomic, lone-pair electron, and σ and π bond regions. These multiple partial charges per atom should better account for the polarization effect than single atomic charge in other PFFs. To evaluate the PFF, structural and energetic properties for some organic and biochemical systems, including rotational barriers; binding energies of base pairs; a base−base interaction in a B-DNA decamer; and interaction energies of ten stationary conformers of a water dimer, peptides, and bases with water molecules, have been calculated and compared with the experimental data or *ab initio* MP2 results. Molecular dynamics simulations using the PFF have been performed for crambin and BPTI protein systems. Better performances in modeling root-mean-square deviations of backbone bond lengths, bond angles, key dihedral angles, the coordinate root-mean-square shift of atoms, and the distribution of hydrogen bonds have been observed in comparison with other PFFs. These results indicate that the fluctuating charge force field, ABEEM$\sigma\pi$/MM, is accurate and reliable and can be applied to wide ranges of organic and biomolecular systems.

Introduction

Molecular dynamics (MD) simulations using force fields (FF) are still an important tool in understanding structure, dynamics, and function properties for biological systems. Although quantum calculations and *ab initio* MD modeling have tremendously advanced in recent decades, they are limited to relatively small systems.[1,2] In a primitive force field, molecules are represented by a collection of atom-centered interaction sites with fixed partial charges. The electrostatic energy is determined by the Coulombic interaction between the partial charges without any inclusion of the polarization effect. In view of the significance of the polarization effect, much has been achieved since the 1970s in the development and employment of the polarizable force field (PFF) for biomolecule simulation, such as CHARMM,[3−8] OPLS,[9−16] AMBER,[17−20] NEMO,[21] AMOEBA,[22,23] QMPFF,[24−26] and so forth. The first physically consistent microscopic study of dielectric effects in nonpolar environments was reported by Warshel and Levitt.[27] Accounts and reviews of PFF are available elsewhere.[27,28] Two approaches have been employed to account for the polarization effect. One is through the induced dipole or multipole (including Drude charge oscillator and induced dipole mixed fluctuating charge models), and the other is to use fluctuating partial charge.

In the induced dipole (multipole) PFF, the potential energy function is augmented by an inductive term from the induced dipole. The contribution of the electrostatic interaction comes from both permanent charges and induced dipole moments obtained from the atomic polarizabilities through an iterative procedure.[29] Karlström et al.[21] have demonstrated the

* Corresponding author e-mail: zzyang@lnnu.edu.cn.
[†] Liaoning Normal University.
[‡] University of North Carolina.

10.1021/ct9006647
Published on Web 02/04/2010

importance of the quadrupole moment. Higher order multipoles were included in the AMOEBA PFF.[22] A new PFF of this category, the X-POL potential,[30] was recently proposed by Xie and co-workers. The published CHARMM all-atom parameters for nucleic acids[4] and proteins[5] provide a consistent set for condensed-phase simulations of a wide variety of biological molecules. Xie et al. developed a polarizable intermolecular function for liquid amides and alkanes by employing the "standard" CHARMM force field plus a polarizable term.[8] The OPLS series of force fields have been developed for more than 20 years and have proved to be highly successful in computing liquid state thermodynamic properties[15] and in dipeptide, protein, and protein−ligand modeling.[11−13] The AMBER force field has simulated the structures, conformational energies, and interaction energies of proteins, nucleic acids, and many related organic molecules in condensed phases.[18] Its point-charge all-atom force field for proteins and united-atom force field for simulations involving highly demanding conformational sampling such as protein folding and protein−protein binding have been advanced.[20] The QMPFF[24−26] was fitted solely to QM data at the MP2/aTZ(-hp) level and had demonstrated high accuracy and transferability in crystal and liquid simulations as a result of its strong physical basis and explicit polarizability. As is well-known, there are many other well-developed useful force fields, such as MM1-MM4,[31] MMFF,[32] ECEPP,[33] CFF,[34] Tripos,[35] GROMOS,[36] etc.

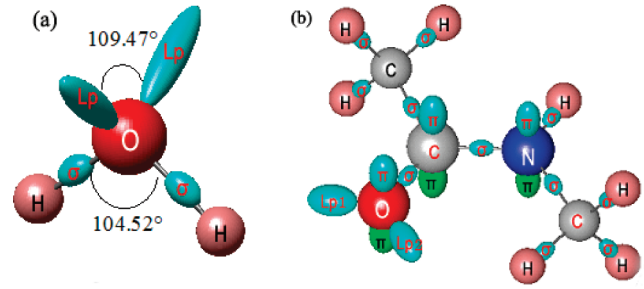
In fluctuating charge models, whose basis is the electronegativity equalization method (EEM) in density functional theory (DFT),[37,38] atomic partial charges of a molecular system are allowed to change with geometry and ambient environment. In the usual EEM scheme, using the atomic partial charge and two characteristic parameters per atom, the electrostatic energy is written as

$$E_{es} = \sum_a (\chi_a^* q_a + \eta_a^* q_a^2) + \sum_{a<b} k_{ab} \frac{q_a q_b}{R_{ab}} \qquad (1)$$

where q_a is the atomic partial charge at the a region (atom and/or specified one), χ_a^* and η_a^* are the valence-state electronegativity and valence-state hardness of region a, R_{ab} is the separation between regions a and b, and the summation for both a and b is over all sites. In DFT, the effective electronegativity χ_a of a site a is equal to the partial derivative of the electrostatic energy with respect to the partial charge q_a of site a. k_{ab} is a correction factor of the Coulombic interaction energy between the partial charges q_a and q_b, which stems from the reality that q_a and q_b involve the electron clouds rather than the ideal point charges.[39−42] According to the electronegativity equalization principle based on DFT, at the equilibrium state, the effective electronegativities of all sites are equal to the global molecular electronegativity χ_{mol}, which constitutes the electronegativity equalization equation. By solving these equations with the charge constraint for a molecular system, the partial charges of all sites in the system are directly and quickly obtained.

There are several fine implementations of EEM to allow rapid calculations of the partial charge distribution in molecules. Mortier et al. proposed a systematic formulation[43] to directly calculate atomic partial charges of a large molecular system. York and Yang[44] established a chemical potential equalization model. Rappé and Goddard presented the charge equilibration method[39] for MD simulations. Stern et al. formulated a combined fluctuating charge and polarizable dipole force field for water, amino acids, and peptides based on *ab initio* data.[45] Recently, Chelli and Procacci developed a transferable polarizable electrostatic force field[46] by generalizing Mortier's method to include atom-based dipolar distributions as done in the chemical potential equalization model by York and Yang. A CHARMM fluctuating charge force field[7] for proteins has been applied to MD simulations for bulk organic liquid and small proteins. Notice that, in all the existing methods, only one partial charge per atomic site is employed.

Chart 1. The Sketch of All Regions in (a) Water and (b) NMA Molecules Defined in the ABEEM$\sigma\pi$ Model



As recently pointed out by Jorgensen, broadly applicable PFFs have yet to emerge.[47] The development of PFFs remains to be a frontier challenge in molecular modeling. Key to the development of a reliable PFF is to effectively account for the polarization effect. In this work, we present a PFF with multiple fluctuating partial charges per atom, ABEEM$\sigma\pi$/MM, based on the atom-bond electronegativity equalization method (ABEEM) that we have recently developed.[48−54]

Model

ABEEM$\sigma\pi$ with Multiple Partial Charges per Atom. Conventional EEM models partition a molecular system into individual atomic regions, each having one partial charge only. As is well-known, besides nuclei, electron densities also concentrate around chemical bonds and lone-pair regions to some extent. Henceforth, to better account for the polarization effect, we suggested that a molecular system is decomposed into atomic regions, lone-pair regions, and σ and π bond regions, and each region is assigned by a partial charge.[46,49,51] Better elucidation of the polarization effect from this partition scheme comes from the greater freedom and larger flexibility in calculating the fluctuating partial charge associated with each of the regions. In principle, if the total number of charge sites approaches infinity, the partial charge q_i will resemble the continuous charge density, the fundamental variable of DFT. This is the essence and foundation of our ABEEM$\sigma\pi$ model.

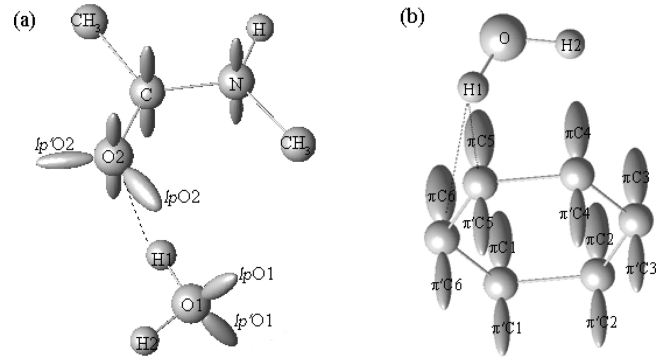
In Chart 1, as examples, we draw the charts of all regions in water and N-methylacetamide (NMA) molecules in the

ABEEM$\sigma\pi$ model. In addition to atomic sites, some new virtual sites including lone pair and σ and π bond sites are placed according to the physical meaning. A water molecule contains seven electron cloud regions in the ABEEM$\sigma\pi$ model: three atoms, two σ bonds whose angle is 104.5°, and two lone pairs whose angle is 109.47°. As is known, there are four electron pairs around the oxygen atom, which are spread so as to point roughly toward the apexes of a tetrahedron. Every atomic charge is placed in the position of the corresponding atom. The σ bond charge is assumed to locate on the point that partitions the bond length according to the ratio of covalent radii of two bonded atoms, and the lone pair sites are placed on the points which are 0.74 Å from the oxygen nucleus with an intervening angle of 109.47° between two lone pairs on oxygen atom. Thus, there are seven partial charge sites for a water molecule in the ABEEM$\sigma\pi$ model.

Consider the NMA molecule, as shown in Chart 1b. The C atom of carbonyl connects two single bonds and one double bond. The geometry around this C atom is trigonal planar. The geometry around the O atom is trigonal planar too, because the O atom of carbonyl connects one double bond and two lone pairs. All the angles between the two lone pairs and C═O bond are 120°. The oxygen atom involves six partial charges, namely, one centered on the oxygen nucleus, one σ region, two π separate upper and lower regions, and two lone pairs. The σ bond partial charge shared by the oxygen and carbon atoms is on the bond at the point that partitions the bond length according to the ratio of covalent radii between O and C atoms; the π bond partial charges are placed above and below the O atom at the covalent radius (0.74 Å) of the O atom perpendicular to the plane formed by the σ bonds and may have different values depending on the environment; the two lone pair partial charges are placed in the covalent radius of the oxygen atom (0.74 Å). An electron pair of the N atom can be used to make a delocalized π bond with carbonyl. There are also similar π bond partial charges for the nitrogen atom and carbon atom of carbonyl in NMA. So the nitrogen atom involves six partial charges, including one atom and three σ and two π regions. As a whole, besides 12 atomic sites, a NMA molecule has an additional 19 sites: 11 σ bond sites, 6 π bond sites, and 2 lone pair sites.

For a molecule, the electrostatic energy is written as eq 1 in the ABEEM$\sigma\pi$ model which contains the regions or sites of the atoms, bonds, and lone pairs, as shown in Chart 1. On the basis of DFT, the effective electronegativity χ_a of every site a can be also expressed as the partial derivative of the electrostatic energy with respect to the partial charge q_a of site a. According to the EEM, at the equilibrium state the effective electronegativities of all sites, including atoms, bonds, and lone pairs are equal to the global molecular electronegativity χ_{mol} for every molecule, which constitutes the electronegativity equalization equation, i.e., $\chi_a = \chi_b = \cdots = \chi_{\mathrm{mol}}$. It can be shown that the number of the equations of EEM is equal to the number of the sites of molecules. These equations, together with the charge constraint and given parameters (valence-state electronegativity χ_a^* and valence-state hardness η_a^* of region a), can be explicitly and quickly solved to give the global molecular electronegativity χ_{mol} and the partial charge q_i on each site i. The detailed formulation of ABEEM$\sigma\pi$ is available in the Supporting Information (SI1).

Chart 2. Sketch of (a) the Cluster NMA+H_2O and (b) the Cluster Benzene+H_2O, Where Hydrogen Atoms in Aromatic Ring and σ Bond Sites Are Omitted



In NMA, the atomic partial charge plus four or five negative partial charges around a non-hydrogen atom, like the C, N, or O atom, may fluctuate, giving a proper response to the geometry and the environmental change. For example, charges of *lp*O2 and *lp′*O2 of NMA are −0.1689 and −0.1639, respectively. In cluster NMA+H_2O (Chart 2a), the charges located on *lp* of O2 all become more negative than that on *lp* electrons of O2 in NMA due to the intermolecular HB. Because the H1 points to the *lp*O2, the charge located on *lp*O2 (−0.1748) is more negative than that on *lp′*O2 (−0.1691). All the π bond charges of all carbon atoms in benzene are −0.0135. In cluster benzene+H_2O, as shown in Chart 2b, the largest polarized regions are πC5 and πC6 partial charge regions, whose charges are −0.0141. All the charges on πC1∼πC4 partial charge regions, which are slightly polarized, are −0.0139; the charges on π'C1∼π'C6 partial charge regions are all −0.0134.

ABEEM$\sigma\pi$ Polarizable Force Field, ABEEM$\sigma\pi$/MM. The energy function $E_{\mathrm{ABEEM}\sigma\pi}$ of the ABEEM$\sigma\pi$ polarizable force field can be written as the sum of following terms:

$$E_{\mathrm{ABEEM}\sigma\pi} = E_{\mathrm{b}} + E_\theta + E_\phi + E_{\mathrm{imptors}} + E_{\mathrm{vdw}} + E_{\mathrm{elec}} \tag{2}$$

where E_{b} and E_θ are the usual energy terms from bond stretching and angle bending contributions modeled as harmonic oscillators (or Morse function for water), respectively,

$$E_{\mathrm{b}}(r) = \sum_{\mathrm{bonds}} k_r(r - r_{\mathrm{eq}})^2 \tag{3}$$

$$E_\theta(\theta) = \sum_{\mathrm{angles}} k_\theta(\theta - \theta_{\mathrm{eq}})^2 \tag{4}$$

where k_r and k_θ represent the force constants of the stretching and bending, r and θ are the actual values of the bond length and bond angle, and r_{eq} and θ_{eq} denote the equilibrium values of the bond length and bond angle, respectively. E_ϕ is the torsional energy for a bond rotation. The torsional term takes the following form:

$$E_{\phi}(\phi) = \sum_{\text{torsions}} \left[\frac{V_1}{2}(1 + \cos\phi) + \frac{V_2}{2}(1 - \cos 2\phi) + \frac{V_3}{2}(1 + \cos 3\phi) \right] \quad (5)$$

and the improper dihedral angle term is written as

$$E_{\text{imptors}} = \sum_{\text{imptors}} v(1 - \cos 2\phi) \quad (6)$$

where V_1, V_2, V_3, and v are the dihedral angle and improper dihedral angle force constants, respectively. E_{vdw} describes the van der Waals nonbonded atom−atom interaction:

$$E_{\text{vdw}} = \sum_{i<j} 4 f_{ij} \varepsilon_{ij} (\sigma_{ij}^{12}/r_{ij}^{12} - \sigma_{ij}^{6}/r_{ij}^{6}) \quad (7)$$

Geometric combining rules for the Lennard-Jones coefficients are employed: $\sigma_{ij} = (\sigma_{ii}\sigma_{jj})^{1/2}$ and $\varepsilon_{ij} = (\varepsilon_{ii}\varepsilon_{jj})^{1/2}$. The summation runs over all of the pairs of unbonded atoms i and j. If i and j are intramolecular, the coefficient $f_{ij} = 0.0$ for any $i-j$ pair connected by a valence bond (1−2 pairs) or a valence bond angle (1−3 pairs), $f_{ij} = 0.5$ for 1,4 interactions (atoms separated by three bonds), and $f_{ij} = 1.0$ for all other intramolecular and intermolecular cases.

The key difference between the ABEEM$\sigma\pi$ PFF and other force fields is the treatment of the electrostatic interaction energy. The electrostatic interaction energy, E_{elec}, is actually expressed as

$$E_{\text{elec}} = \sum_{i<j} k_{ij} q_i q_j / r_{ij} \quad (8)$$

where r_{ij} is the distance between sites i and j. q_i and q_j are the partial charges of regions or sites i and j, which are calculated by the ABEEM$\sigma\pi$ method. In reality, either q_i or q_j involves or represents some sort of electron clouds. Therefore, when we model them as point charges in EEM or ABEEM, their electrostatic interaction energy is expressed as $k_{ij}q_iq_j/r_{ij}$ rather than the pure Coulombic form q_iq_j/r_{ij}. The introduced parameter k_{ij} may be said to be a result of considering the exchange, penetration, and shielding effect in the interaction between the two pieces of electron clouds i and j. There is a similar parameter, the electrostatic interaction factor (modified Coulomb interaction), $J_{ij}(r_{ij})$, such as in the models of the charge equilibration of Rappé and Goddard,[39] Bakowies and Theil,[40] as well as Dias and co-workers,[41,42] which was taken into account in their electrostatic energy expressions and is related to the local hardnesses of atoms i and j.

In our ABEEM or MEEM model, k_{ij} is optimized and set to be 0.57 empirically, except for hydrogen-bond regions. In the hydrogen bond interaction region,[51,53] k_{ij} is replaced by a $k_{\text{HB}}(r_{ij})$ function to describe the electrostatic interaction between the hydrogen atom and the lone-pair electron.

In molecular simulations, we use the ABEEM$\sigma\pi$ method to calculate partial charges of all regions, namely, atoms, lone pairs, σ bonds, and π bonds, and then employ eq 2 to compute the total energy of the system. If there is a geometrical change, i.e., a change of bond length, angle, dihedral angle, or relative position between molecules, we will recalculate the partial charges using the ABEEM$\sigma\pi$ method, and then the total energy. In this manner, a systematic way to account for the polarization effect by allowing partial charge fluctuations in accordance with the changing molecular environment is provided by the ABEEM$\sigma\pi$ force field, termed in short as ABEEM$\sigma\pi$/MM.

ABEEM$\sigma\pi$/MM Parameters' Determination and Calibration. The systematic determination and calibration of parameters in developing any force field is tedious, involving tremendous amount of tasks in testing, calibrations, and analyses. In our case, more than 2000 organic and biological molecules were chosen. A large amount of efforts have been invested to optimize the ABEEM$\sigma\pi$/MM parameters to make them consistent, reliable, and transferable in accurately reproducing structural, energetic, and dynamic properties.

As is known, FF involves a series of parameters, such as bond stretching parameters, angle bending parameters, torsional parameters, van der Waals parameters, as well as the valence-state electronegativities and valence-state hardnesses in the electrostatic energy term etc. Strenuous and time-consuming work is needed for calibration of the parameters for development of a good FF.

The parameters χ^* and η^* are fitted through a regression and least-squares optimization procedure and are listed in Table S1 (Supporting Information), where the scaling Pauling electronegativity unit is used. The ABEEM$\sigma\pi$ labels are listed in Chart S1 (Supporting Information). The parameters are fitted not only to reproduce the charges of the *ab initio* calculation but also to fit the dipole moments, structures, and dimer binding energies given by experimental measurements and *ab initio* calculations. The fitting function k_{HB} is extracted from k in eq 8 to describe the electrostatic interaction between the H atom and *lp* of the acceptor. A general formulation of the k_{HB} function is available in SI1 (Supporting Information).

It is well-known that parameters for the hard degrees of freedom (bond stretching and angle bending) can be transferred from one FF to another without modification. For example, for some bonds and angles in the amide system, the same force constants, equilibrium bond lengths, and bond angles are used in both OPLS-AA[10] and AMBER[18] FFs. So we take the parameters of bond stretching and angle bending of protein directly from OPLS-AA FF and those of DNA from AMBER FF. The torsional terms are often regarded as "soft" degrees of freedom, in which most of the variations in structure and relative energy are due to the complex interplay between the torsional and nonbonded contributions. In ABEEM$\sigma\pi$ FF, we take the torsional and improper torsional parameters of OPLS-AA or AMBER as a reference and refit them through the least-squares optimization procedure to make the conformation energies and the key dihedral angles' root of mean square deviation (rmsd) of the model molecules be in good agreement with those calculated by the *ab initio* method. In addition, the Lennard-Jones parameters are determined by fitting *ab initio* conformational energies, dimer binding energies, dipole moments, and so on, using a regression and least-squares method. All the parameters are summarized in Table S1 (Supporting Information).

It should be pointed out that the parameters we have used are transferable over a large amount of chemical and

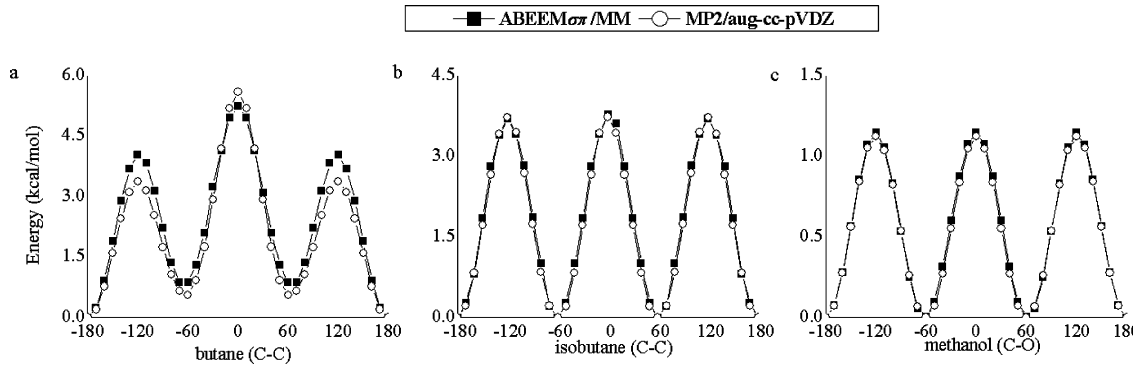


Figure 1. Rotational energy profiles (a) for the central C−C bond rotation of butane, (b) for the C−C bond rotation of isobutane, and (c) for the C−O rotation of methanol from MP2/aug-cc-pVDZ calculations and the ABEEM$\sigma\pi$ force field.

biological species rather than needing to be reparameterized for each molecule. The charge parameters, such as valence electronegativity and hardness for a type of site or region, are used for all molecules rather than only for one molecule in the ABEEM$\sigma\pi$ model. A site type in the ABEEM$\sigma\pi$ model is properly defined according to the type of its surrounding chemical environment, so that the parameters are transferable, only depending on the site types as in the usual force fields.

For example, a single water molecule has 7 total sites: three atomic sites, two σ bond sites, and two lone pair sites. Each site has two charge parameters (one valence electronegativity and one valence hardness). Besides the charge parameters, a single water molecule has 9 other parameters which contain r_{OH}, θ_{HOH}, k_{HOH}, ε_O, ε_H, σ_O, and σ_H, as well as α and D, because the stretch vibration of every O−H bond is described by the Morse potential function.

The ABEEM$\sigma\pi$/MM parameters and their calibration results are available in SI1 (Supporting Information).

Computational Details of Molecular Dynamics. The structures of Crambin, BPTI, and Trypsin from the protein data bank were used as the initial geometries of MD simulations. The MD simulations were performed using the modified TINKER program in the NVT ensemble with Berendsen thermostats, the velocity Verlet integrator, and a time step of 1 fs. The systems were initially heated over 5 ps to 285 K. The cutoff radius for nonbonding interactions was 10.0 Å with the minimum image convention if the periodic boundary condition was used. For all simulations, 0.5 ns of a MD run for equilibration was performed, followed by 9.5 ns of simulations for the calculation of various properties. We recomputed the partial charges of all sites using the ABEEM$\sigma\pi$ method every 0.1 ps. The charges are placed on some virtual sites according to the physical meaning, including the sites of σ and π bonds and lone pairs, for better calculating the electrostatic energy. When the MD simulation was performed, the force was only acting on the atoms with redistribution of the partial charges of bonds and lone pairs to the connected atoms. The partial charge of one σ bond is averagely assigned to two bonded atoms. The charges of π regions and lone pairs are all assigned to the connected atoms.

We performed simulations of 1crn by ABEEM$\sigma\pi$ (1792 sites) and standard AMBER (642 sites) force fields, respectively. These simulations have a time step of 1 fs, a simulation time of 1 ns, and a cutoff radius of 10.0 Å. With the same computational power, the AMBER fixed charge FF expends 10851.52 s. If the partial charges were fixed, the ABEEM$\sigma\pi$ FF expends 12093.66 s, and if the partial charges were recomputed every 0.1 ps, ABEEM$\sigma\pi$ FF uses 149116.20 s.

In what follows, we present results for representative systems obtained by the ABEEM$\sigma\pi$/MM and compare them with experimental data or results from the *ab initio* method at the MP2 level.

Results and Discussion

Rotational Barriers. Rotational energy profiles for the central C−C bond of butane and isobutane, and for the C−O rotation of methanol obtained at the ABEEM$\sigma\pi$/MM and MP2/aug-cc-pVDZ levels of theory, are shown in Figure 1. One can see that ABEEM$\sigma\pi$/MM accurately reproduces the rotational energy profiles. The experimental syn rotation and trans−gauche barriers of butane are 4.56[55] and 3.30 kcal/mol,[56] respectively. The corresponding results from ABEEM$\sigma\pi$/MM are 4.38 and 3.18 kcal/mol, whereas MP2 gives 4.95 and 2.72 kcal/mol, and AMBER[57] FF yields 5.31 and 3.53 kcal/mol, respectively. The experimental energy difference between trans and gauche isomers is 0.497−0.89 kcal/mol.[58] The result of ABEEM$\sigma\pi$/MM, MP2, and AMBER FF is 0.86, 0.65, and 0.79 kcal/mol, respectively. For isobutene, ABEEM$\sigma\pi$/MM yields 3.78 kcal/mol, comparable to the experimental[59] result of 3.90 kcal/mol and the AMBER[57] and MP2/aug-cc-pVDZ results, which are 3.48 kcal/mol and 3.61 kcal/mol, respectively. The experimental[60] barrier height of methanol is 1.07 kcal/mol, and the corresponding value from ABEEM$\sigma\pi$/MM and MP2/aug-cc-pVDZ is 1.12 and 1.14 kcal/mol, better than the results from MMFF94[61] and MM3[62] FF, which are 1.23 and 0.78 kcal/mol, respectively.

Structures of Peptides. Alanine dipeptide is a model compound of the protein backbone, containing two peptide linkages. We use the *ab initio* result at the level of MP2/6-311++G(d,p)//HF/6-31G(d,p) as the reference. The same as for the *ab initio* result, ABEEM$\sigma\pi$/MM is able to find all six of its stable conformers, especially β_2 and α_L, which are difficult to locate because they are shallow minima. Table S3 (Supporting Information) shows conformational energies

Table 1. Conformational Energies (in kcal/mol), RMSD of Conformational Energies, and the Key Dihedral Angles (in deg) for Alanine Tetrapeptide Relative to the *ab Initio* Data[a]

	ab initio[b]	ABEEM$\sigma\pi$	*ab initio*[c]	ABEEM[d]	OPLS-AA/L[e]	FQ-Dipole[f]	OPLS-FQ[g]	AMBER[h]	CHARMM-FQ[i]
1	4.36	4.85/2.8	2.71	2.88/5.6	3.19/4.4	2.88/-	3.03/-	3.52/-	5.15/-
2	3.76	2.54/3.5	2.84	1.94/8.3	3.19/6.5	1.84/-	3.97/-	3.74/-	3.42/-
3	0.00	0.00/3.7	0.00	0.00/7.8	−0.32/8.4	0.22/-	0.26/-	0.00/-	−0.70/-
4	4.67	4.22/5.6	4.13	2.90/8.2	4.40/5.8	3.69/-	2.34/-	4.32/-	5.36/-
5	4.64	4.15/5.1	3.88	4.81/7.3	3.14/9.3	3.70/-	4.62/-	3.64/-	4.21/-
6	0.66	0.67/4.0	2.20	1.80/8.8	0.96/12.7	1.45/-	1.67/-	2.91/-	2.40/-
7	4.13	4.51/9.0	5.77	5.83/11.7	5.82/6.6	5.48/-	6.18/-	4.54/-	5.75/-
8	4.16	4.16/12.7	4.16	4.10/12.7	4.83/18.8	5.38/-	4.39/-	5.91/-	3.40/-
9	6.13	5.76/3.4	6.92	7.21/5.9	7.14/8.2	6.74/-	5.33/-	4.93/-	6.10/-
10	5.10	4.99/6.0	6.99	8.03/4.3	7.25/14.2	8.21/-	7.82/-	8.53/-	4.55/-
rmsd		0.50/6.3[j]		0.67/8.4[k]	0.56/10.4[k]	0.71[k]	0.94[k]	1.14[k]	1.25[k]

[a] The values before "/" are the conformational energies, and those after "/" are the rmsd of the key dihedral angles relative to HF/6-31G(d,p) level geometries, which are listed in Table S2 (Supporting Information). The FQ-dipole, OPLS-FQ, AMBER, and CHARM-FQ methods did not provide the definite dihedral information. [b] The conformational energies at the MP2/6-311++G(d,p)//HF/6-31G(d,p) level calculated in this work. [c] The conformational energies at the LMP2/cc-pVTZ(-f)//HF/6-31G(d,p) level are cited from ref 5. [d] Ref 53, ABEEM FF only involves the partial charge sites of atoms and bonds without adding lone pair sites and π charge sites. [e] Ref 12. [f] Ref 45. [g] Ref 11. [h] Ref 19. [i] Ref 7. [j] The lowest rmsd relative to the results of the MP2/6-311++G(d,p) calculation. [k] The lowest rmsd relative to the results of the LMP2/cc-pVTZ(-f) calculation.

and rmsd's of key dihedral angles for alanine dipeptide from ABEEM$\sigma\pi$/MM, ABEEM/MM,[53] OPLS-AA/L,[12] OPLS-PFF,[13] FQ-dipole,[45] and OPLS-FQ[11] methods in comparison with *ab initio* results. It is seen from the table that the ABEEM$\sigma\pi$/MM method best reproduces both conformational energies and the dihedral angles for alanine dipeptide, giving the lowest rmsd for both quantities. Its rmsd value relative to the *ab initio* MP2/cc-pVTZ//MP2/6-31G(d,p) result[63] in conformational energies is merely 0.20 kcal/mol, and that of the key dihedral angles is only 5.9°. In comparison with ABEEM/MM, whose only difference from ABEEM$\sigma\pi$/MM is without partial charge sites for lone pairs and π bonds, a marked difference in rmsd values is also apparent.

For alanine tetrapeptide, Table 1 displays a rmsd in conformational energies and key dihedral angles obtained from ABEEM$\sigma\pi$/MM and other methods, compared to *ab initio* results. ABEEM$\sigma\pi$/MM gives a rmsd of 0.50 kcal/mol in conformational energies and 6.3° in the key dihedral angles relative to the MP2/6-311++G(d,p)//HF/6-31G(d,p) data, whereas those quantities are 0.67 kcal/mol and 8.4° for ABEEM/MM,[53] 0.56 kcal/mol and 10.4° for OPLS-AA/L,[12] and 0.69 kcal/mol and 19.1° for OPLS/PFF,[13] respectively. For FQ-Dipole,[45] OPLS-FQ,[11] AMBER,[19] and CHARMM-FQ,[7] the rmsd in conformational energies is 0.71, 0.94, 1.14, and 1.25 kcal/mol, respectively, in comparison with the LMP2/cc-pVTZ(-f)//HF/6-31G(d,p) result. No dihedral angle rmsd result is available for those latter approaches for this system.

Table 2 summarizes rmsd values in conformational energies and key dihedral angles obtained for the ABEEM$\sigma\pi$/MM, OPLS-AA/L, and OPLS/PFF methods in comparison with the *ab initio* result for various neutral dipeptides and tetrapeptide. The average rmsd in conformational energies and key dihedral angles of neutral peptides are 0.36 kcal/mol and 4.3° for ABEEM$\sigma\pi$/MM, respectively, but 0.47 kcal/mol and 10.1° for OPLS-AA/L[12] and 0.43 kcal/mol and 10.5° for OPLS/PFF, respectively.[13] For phenylalanine, tryptophan, tyrosine, and histidine dipeptides containing at least one aromatic ring, because ABEEM$\sigma\pi$/MM explicitly adds partial charges to π bonds, significantly lower rmsd values in angles are obtained, 3.4° on average for the four dipeptides from ABEEM$\sigma\pi$/MM, whereas the corresponding result from OPLS-AA/L[12] and OPLS/PFF[13] is 14.6° and 14.0°, respectively. Hydrogen bonding is closely related to the orientation of lone pairs of polar atoms. For asparagine and glutamine dipeptides with more hydrogen bonds, since ABEEM$\sigma\pi$/MM explicitly augments partial charges sites for lone pairs, noticeable improvements in rmsd are observed, 0.00 and 0.85 kcal/mol in conformation energies and 6.8° and 4.8° in dihedral angles for the two peptides, respectively, compared to 0.16 and 0.96 kcal/mol and 19.5° and 13.9° from OPLS-AA/L[12] and 0.02 and 0.92 kcal/mol and 8.7° and 18.0° from OPLS/PFF[13] for them. ABEEM$\sigma\pi$/MM results agree well with OPLS-AA/L and OPLS/PFF results for aliphatic amino acid dipeptides, such as valine, leucine,

Table 2. RMSDs in Conformational Energies (in kcal/mol) and the Key Dihedral Angles (in deg) for Different Peptides Relative to the *ab Initio* Result

molecule	ABEEM$\sigma\pi$	OPLS-AA/L[a]	OPLS/PFF[b]
Di-Ala	0.20/5.9	0.27/6.5	0.35/7.1
Tetra-Ala	0.50/6.3	0.56/10.4	0.69/19.1
Di-Phe	0.00/2.7	0.15/7.5	0.02/9.5
Di-Trp	0.68/4.2	0.50/24.2	0.49/19.4
Di-Tyr	0.27/3.4	0.39/8.1	0.27/8.9
Di-His	1.01/3.5	0.85/18.7	0.83/18.2
Di-Asn	0.00/6.8	0.16/19.5	0.02/8.7
Di-Gln	0.85/4.8	0.96/13.9	0.92/18.0
Di-Val	0.00/1.9	0.08/8.4	0.01/5.1
Di-Leu	0.30/3.0	0.34/6.1	0.35/5.1
Di-Ile	0.59/5.3	0.38/5.5	0.88/11.8
Di-Ser	0.33/5.4	0.44/4.9	0.34/8.1
Di-Cys	0.11/3.3	0.35/5.8	0.27/4.8
Di-Met	0.24/3.2	0.59/5.2	0.53/5.4
Di-Thr	0.37/5.3	0.87/7.1	0.75/8.9
average	0.36/4.3	0.47/10.1	0.43/10.5
Di-Asp	0.15/4.1	0.16	0.77
Di-Glu	1.29/5.0	1.53	1.47
Di-Lys	0.58/2.9	0.88	0.59
Di-Pro his	0.70/5.9	0.97	0.97
Di-Arg	1.14/4.1	1.15	0.79
average	0.77/4.4	0.94	0.92

[a] Ref 12. [b] Ref 13.

Table 3. RMSD and Linear Regression Analyses of Interaction Energies for H-Bonded Nucleic Acid Bases Relative to MP2/6-31G(d)(0.25) Values

force field	rmsd	R^a	SD^b	A^c	B^c	AAD^d
ABEEM$\sigma\pi$	0.90	0.98	0.77	−0.87	0.97	0.7
CHARMM27[e]	1.16	0.96	1.15	−0.34	0.96	0.9
AMBER4.1[f]	1.21	0.98	0.93	−0.59	1.10	0.9
CVFF[f]	4.83	0.88	1.30	0.79	0.62	4.4
CFF95[f]	1.72	0.95	1.05	1.64	0.80	1.2
OPLS[f]	2.65	0.95	1.76	−1.69	0.95	2.4

[a] Correlation coefficient. [b] Standard deviation (in kcal/mol). [c] The equation is $Y = A + Bx$, where A is intercept, B is slope, and x represents the MP2/6-31G(d)(0.25) values. [d] Average absolute deviation (in kcal/mol). [e] Ref 65. [f] Ref 64.

and isoleucine dipeptides, and polar amino acid dipeptides, including serine, threonine, cysteine, and methionine dipeptides. There is nevertheless an important difference, where OPLS-AA/L and OPLS/PFF performed structure optimizations with all the key dihedral angles constrained to their *ab initio* structure positions for the charged dipeptides, but in ABEEM$\sigma\pi$/MM calculations no such constraints were applied. The average rmsd value in conformational energy was 0.77 kcal/mol for ABEEM$\sigma\pi$/MM, whereas that quantity was 0.94 and 0.92 kcal/mol for OPLS-AA/L and OPLS/PFF, respectively. Furthermore, ABEEM$\sigma\pi$/MM obtained a rather small average angular rmsd of 4.4°. Proline is a special case, where disulfide bridges stabilize polypeptides and proteins. We considered several conformations for them whose calculated RMSDs were also found to be small, as listed in Table S3 (Supporting Information).

Interactions in Base Pairs and Water Dimer, Dipeptide−Water and Base−Water Complexes. We computed the interaction energies with ABEEM$\sigma\pi$/MM and compared them with other FFs and MP2/6-31G(d)(0.25) results of 26 H-bonded base pairs from Hobza et al.[64] Table 3 exhibits their rmsd's and linear regression analyses. The rmsd of ABEEM$\sigma\pi$/MM is 0.90 kcal/mol, smaller than that of CHARMM27,[65] AMBER4.1,[64] CFF95,[64] and OPLS[64] force fields. The correlation coefficient R is 0.98, and the standard deviation is 0.77 kcal/mol, with the intercept A and slope B in the fitted linear function of $Y = A + Bx$ being −0.87 and 0.97, respectively. The average absolute deviation is 0.7 kcal/mol, which is also the smallest among the methods tested. The good agreement between ABEEM$\sigma\pi$/MM results and high-level *ab initio* data indicates that ABEEM$\sigma\pi$/MM can correctly predict the interaction energy between H-bonded nucleic acid bases. The detailed information of the above calculations is contained in Table S4 (Supporting Information).

The stacked structures of A and T (denoted as ATs) and C and G (denoted as CGs) base pairs are what we investigated next. Their initial geometries (shown in Figure 2) are from ref 66. We calculated interaction energies of the stacked bases as a function of the distance between two planes formed by two base pairs. Single point calculations were done with MP2/6-31G(d), MP2/6-311++G(d,p), and ABEEM$\sigma\pi$/MM methods. The vertical separation between A and T base planes and between C and G planes was varied from 2.8 Å to 10.0 Å. The potential energy profiles of ATs and CGs are shown in Figure 2a and b, respectively. One

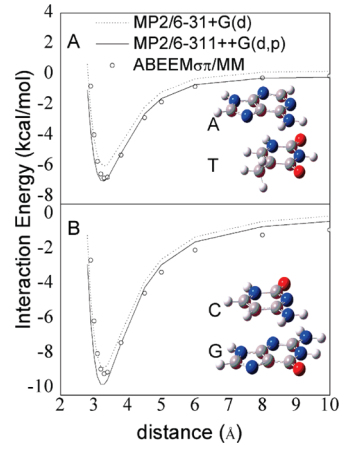


Figure 2. Potential energy profiles of stacked nucleic acid base pairs. (a) ATs. (b) CGs.

Table 4. Lowest Interaction Energies of Stacked Nucleic Acid Base Pairs (kcal/mol)

method	ATs	CGs
MP2/6-31+G(d)	−6.0	−8.9
MP2/6-311++G(d,p)	−7.0	−9.9
ABEEM$\sigma\pi$/MM	−6.8	−9.3
PMP[a]	−9.7	−10.9
AMBER[a]	−8.7	−13.0

[a] Ref 66.

local minimum is found with a vertical separation of 3.3 Å for each profile. Table 4 summarizes the lowest interaction energies of ATs and CGs from ABEEM$\sigma\pi$/MM, Nakagawa's polarizable model potential function,[66] AMBER[66] force field, MP2/6-31+G(d), and MP2/6-311++G(d,p) calculations. The interaction energies by Nakagawa's model and AMBER are underestimated. The lowest ABEEM$\sigma\pi$/MM interaction energies for ATs and CGs are −6.8 and −9.3 kcal/mol, respectively, in good agreement with *ab intio* results of −7.0 and −9.9 kcal/mol at the MP2/6-311++G(d,p) level of theory, indicating that ABEEM$\sigma\pi$/MM is suitable to predict the interaction energy for stacked nucleic acid base pairs.

We next considered the monoclinic B-DNA decamer $(CCAACGTTGG)_2$, which has five crystallographically different base pairs. Only nucleic acid bases are considered, and sugar−phosphate units are omitted. The interaction energy for H-bonded pairs, intrastrand stacking pairs, and interstrand stacking pairs from the MP2/6-31G(d) (0.25) (BSSE corrected) calculation, ABEEM$\sigma\pi$, CHARMM27,[65] and AMBER force fields[64] are reported in Table S5 (Supporting Information). At the bottom of Table S5 is the sum of different types of interacting pairs, including 10 H-bonded interactions (ΣH), 18 intrastrand stacking interactions (ΣS), 18 interstrand stacking interactions (ΣI), and the total interaction energies (ΣH + S + I). The sum of H-bonded base pair interaction energy from the ABEEM$\sigma\pi$ force field is −212.23 kcal/mol, which is very close to the −209.4 kcal/mol from the MP2 calculation. The intrastrand stacking interaction energy from ABEEM$\sigma\pi$/MM is −52.76 kcal/mol, higher than that from the MP2 calculation (−71.0 kcal/mol). For the interstrand stacked base pairs, the dominant contributions of electrostatic and Lennard-Jones term are varied. The MP2 interaction energy of interstrand stacked base pairs is

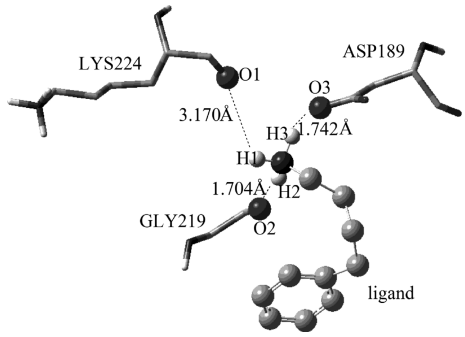


Figure 3. The configuration of the active pocket of 1tni.

well reproduced by ABEEM$\sigma\pi$/MM with a difference of less than 1.2 kcal/mol. The above analyses suggest that there exits a balance between H-binding and stacking interactions in the system, and electrostatic and Lennard-Jones contributions in base−base interactions can satisfactorily be treated by ABEEM$\sigma\pi$/MM.

The treatments of the most stable conformer of water clusters have been given elsewhere.[51] Here, we present the detailed study results for ten water dimer stationary conformers. The ABEEM$\sigma\pi$ force field can obtain all ten lowest conformers of the water dimer and excellently reproduce the structures obtained from counterpoise corrected MP2/aug-cc-pVDZ calculations obtained in this work (Table S6, Supporting Information). The binding energies of the ABEEM$\sigma\pi$ force field are comparable with those results of high level *ab initio* CCSD(T)/6-311++G(3df,3pd) calculations (Table S7, Supporting Information).

We have further investigated other systems of dipeptide−water and base−water clusters, such as (1) alanine dipeptide with $(H_2O)_{1-4}$ by ABEEM$\sigma\pi$/MM and MP2/6-31+G(d)//B3LYP/6-31G(d) methods and (2) bases with $(H_2O)_{1-3}$ by ABEEM$\sigma\pi$/MM and MP2/6-311++G(d,p)//B3LYP/6-311++G(d,p) methods, whose geometries, hydrogen bond energies, and cooperative properties are presented in Table S4 (Supporting Information). Results from all these systems in SI1 (Supporting Information) confirm that reliable predictions of interaction energies can be quantitatively obtained from ABEEM$\sigma\pi$/MM.

Illustration of Partial Charges in a Protein/Ligand Docking System. Protein−ligand recognition is a complicated process, but mandatory for the structure-based drug discovery and design. One quantity that is fundamentally important and essential in the process is to reliably and accurately calculate partial charges for ligands and protein, especially for those atoms near the active site and the interface between the donor and acceptor. For example, to design inhibitors for serine proteinase (PDB ID: 1tni), three hydrogen atoms (H1, H2, and H3) of the amine group at the tail of the ligand are found to form hydrogen bonds and one salt-bridge (between positively charged protonated amine and negatively charged aspartic acid) with three receptor atoms (see Figure 3), O1(LYS224), O2(GLY219), and O3(ASP189). The partial charges on those three hydrogen atoms are 0.415, 0.535, and 0.536 at the *ab initio* HF/6-31G level,[67] respectively. When evaluated by ABEEM$\sigma\pi$/MM, which uses a global scale factor k for the electrostatics energy calculations, their charges are 0.443, 0.519, and 0.566, close to the *ab initio* result and with the same trend. If OPLS-AA is used, each of these three hydrogen atoms has a fixed charge of 0.330. In a recent calculation by Cho et al. using a QM/MM model,[68] their charges are 0.340, 0.360, and 0.440, much smaller than the *ab initio* values. These results reveal that ABEEM$\sigma\pi$/MM not only takes good care of the polarization effect but it can also satisfactorily treat the salt-bridge effect.

MD Simulations for Proteins. To demonstrate the reliability of ABEEM$\sigma\pi$ PFF in molecular dynamics simulations, we performed 10 ns MD simulations for crambin (PDB ID: 1crn) and bovine pancreatic trypsin (PDB ID: 5pti), examined the obtained protein backbone geometries, and then compared them with those from the CHARMM force fields. The average rmsd's of the two proteins in backbone bond lengths are 0.037 Å and 0.040 Å, and the rmsd's in backbone bond angles are 4.16° and 4.73°, respectively. The rmsd value in backbone dihedral angles, φ, Ψ, and ω, from ABEEM$\sigma\pi$/MM is smaller than the result from CHARMM: 25.0°, 23.3°, and 7.6° for crambin and 29.4°, 30.2°, and 7.8° for trypsin from ABEEM$\sigma\pi$/MM and 26.0°, 29.6°, and 6.9° for crambin and 40.8°, 36.2°, and 8.7° for trypsin from CHARMM, respectively. These results demonstrate that ABEEM$\sigma\pi$/MM

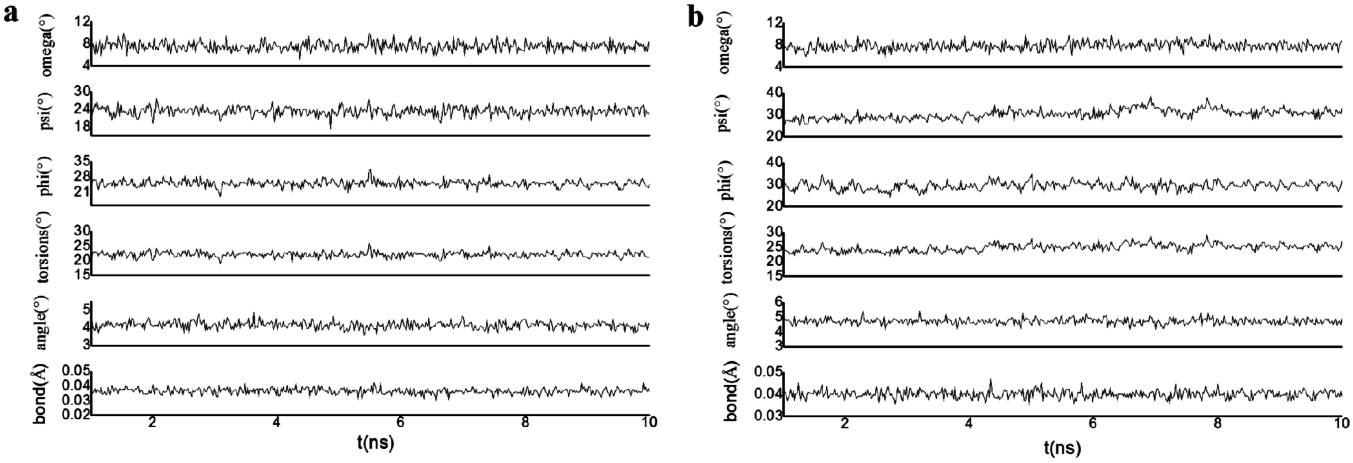


Figure 4. The rmsd in bond lengths, bond angles, and key dihedral angles (φ, Ψ, and ω) between X-ray structures and the ABEEM$\sigma\pi$/MM structures as a function of simulation time for 1crn and 5pti.

can reliably reproduce the protein backbone from the X-ray structure. Figure 4 details the coordinate rmsd of different groups of non-hydrogen atoms, indicating that the structures during the course of MD simulations were stable and the simulations were in the equilibrium state. The rmsd averages over C_α, backbone heavy atoms, all heavy atoms, side chain heavy atoms between experimental structures, and MD simulations using ABEEM$\sigma\pi$ are 2.075, 2.022, 2.222, and 2.502 Å for crambin and 1.786, 1.911, 2.800, and 3.461 Å for trypsin; the rmsd averages over backbone heavy atoms, all heavy atoms, side chain heavy atoms between experimental structures, and MD simulations using CHARMM[5] are 1.700, 1.910, and 2.160 Å for crambin and 2.580, 3.190, and 3.730 Å for trypsin. Average hydrogen bond distances of N···O pairs and O···O pairs of proteins are presented in Table S8 (Supporting Information). ABEEM$\sigma\pi$/MM can accurately describe hydrogen bond interactions for the protein systems. Overall, the results of ABEEM$\sigma\pi$/MM are better than those from CHARMM.[5] This is because, in ABEEM$\sigma\pi$/MM, there is more than one partial charge per atom site, providing more flexibility to account for the polarization effect by allowing different partial charges to adjust their values in accordance with changing geometry and environment during the course of dynamic simulation processes.

Conclusions

A polarizable force field, ABEEM$\sigma\pi$/MM, with multiple fluctuating partial charges per atom site is proposed and evaluated in the present work. The main difference between ABEEM$\sigma\pi$/MM and other force fields is the treatment of the electrostatic interaction energy. In ABEEM$\sigma\pi$ FF, partial charges are evaluated and updated by the ABEEM$\sigma\pi$ method based on the electronegativity equalization principle. The ABEEM$\sigma\pi$ method explicitly considers contributions from lone pairs, σ and π bond regions. This addition increases the flexibility to better account for the polarization effect for both intra- and intermolecular processes.

With high-level *ab initio* [MP2, MP4, and/or CCSD(T)] calculation results or experimental data as the reference, we tested and evaluated ABEEM$\sigma\pi$/MM with a number of biomolecular systems. We found that it can satisfactorily reproduce structural and energetic properties, such as rotational energy profiles, dihedral angles for peptides, conformational energies, partial charge distribution inside an active site, and protein structures in MD simulations. The interaction energies of H-bonded and stacked nucleic base pairs are in good agreement with the available experimental data and *ab initio* results. The ABEEM$\sigma\pi$ polarizable force field also provides reliable binding energies for ten stationary conformers of water dimer, dipeptide−water, and base−water clusters. These results indicate that the fluctuating charge force field, ABEEM$\sigma\pi$/MM, is accurate and reliable and can be applied to wider ranges of biomolecular systems, including aqueous solutions.

Acknowledgment. Helpful comments and suggestions from Professors Robert G. Parr and Lee G. Pedersen of University of North Carolina are gratefully acknowledged. We greatly thank Professor Jay William Ponder for providing the Tinker program; thanks are also given to the support of the National Science Foundation of China (NSFC, Nos. 20633050, 20873055 and 20703022), and the Department of Education of Liaoning Province (Nos. 2008S133, 2009T057, LNET RC0503, and 20060494).

Supporting Information Available: ABEEM$\sigma\pi$ for calculating charge distribution in molecules; ABEEM$\sigma\pi$ force field parameter determination, calibration, and evaluation; training set; interactions in water dimer, dipeptide−water, and base−water complexes. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(2) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *120*, 300–311.

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

(4) MacKerell, A. D., Jr.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.

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

(6) Patel, S.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1–15.

(7) Patel, S.; MacKerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504–1514.

(8) Xie, W.; Pu, J.; MacKerell, A. D., Jr.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1878–1889.

(9) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(10) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(11) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741–754.

(12) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(13) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515–1531.

(14) Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. *J. Comput. Chem.* **1999**, *20*, 704–712.

(15) Jorgensen, W. L.; Severance, D. L. *J. Am. Chem. Soc.* **1990**, *112*, 4768–4774.

(16) Essex, J. W.; Severance, D. L.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **1997**, *101*, 9663–9669.

(17) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(18) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.;

Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(19) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(20) Yang, L. J.; Tan, C.-H.; Hsieh, M.-J.; M., W. J.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollman, P. A.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 13166–13176.

(21) Engkvist, O.; Åstrand, P.-O.; Karlström, G. *Chem. Rev.* **2000**, *100*, 4087–4108.

(22) Kaminský, J.; Jensen, F. *J. Chem. Theory Comput.* **2007**, *3*, 1774–1788.

(23) Penev, E.; Ireta, J.; Shea, J.-E. *J. Phys. Chem. B* **2008**, *112*, 6872–6877.

(24) Donchev, A. G.; Galkin, N. G.; Illarionov, A. A.; Khoruzhii, O. V.; Olevanov, M. A.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8613–8617.

(25) Khoruzhii, O.; Donchev, A. G.; Galkin, N.; Illarionov, A.; Olevanov, M.; Ozrin, V.; Queen, C.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10378–10383.

(26) Donchev, A. G.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, O. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7829–7834.

(27) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(28) Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory Comput.* **2007**, *3*, 2034–2045.

(29) Rasmussen, T. D.; Ren, P.; Ponder, J. W.; Jensen, F. *Int. J. Quantum Chem.* **2007**, *107*, 1390–1395.

(30) Xie, W.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1890–1900.

(31) Wertz, D. H.; Allinger, N. L. *Tetrahedron* **1974**, *30*, 1579–1586.

(32) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.

(33) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361–2381.

(34) Lifson, S.; Hagler, A. T.; Dauber, P. *J. Am. Chem. Soc.* **1979**, *101*, 5111–5121.

(35) Clark, M.; Cramer, R. D., III; van Opdenhosch, N. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(36) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(37) Parr, R. G.; Yang, W. Chemical Potential Derivatives. In *Density Functional Theory of Atom and Molecules*; Oxford University Press and Clarendon Press: New York, 1989; pp 90−95.

(38) Gerrlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1873.

(39) Rappé, A. K.; Goddard, W. A., III. *J. Phys. Chem.* **1991**, *95*, 3358–3363.

(40) Bakowies, D.; Theil, W. *J. Comput. Chem.* **1996**, *17*, 87–108.

(41) Shimizu, K.; Chaimovich, H.; Farah, J. P. S.; Dias, L. G. *J. Phys. Chem. B* **2004**, *108*, 4171–4177.

(42) Menegou, G.; Shimizu, K.; Farah, J. P. S.; Dias, L. G.; Chaimovich, H. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5933–5936.

(43) Mortier, W. J.; Ghosh, S. K.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315–4320.

(44) York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *104*, 159–172.

(45) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730–4737.

(46) Chelli, R.; Procacci, P. *J. Chem. Phys.* **2002**, *117*, 9175–9189.

(47) Jorgensen, W. L. *J. Chem. Theory Comput*, **2007**, *3*, foreword.

(48) Yang, Z. Z.; Cui, B. Q. *J. Chem. Theory Comput.* **2007**, *3*, 1561–1568.

(49) Yang, Z. Z.; Wang, C. S. *J. Phys. Chem. A* **1997**, *101*, 6315–6321.

(50) Cong, Y.; Yang, Z. Z. *Chem. Phys. Lett.* **2000**, *316*, 324–329.

(51) Yang, Z. Z.; Wu, Y.; Zhao, D. X. *J. Chem. Phys.* **2004**, *120*, 2541–2557.

(52) Wu, Y.; Yang, Z. Z. *J. Phys. Chem. A* **2004**, *108*, 7563–7576.

(53) Yang, Z. Z.; Zhang, Q. *J. Comput. Chem.* **2006**, *27*, 1–10.

(54) Yang, Z. Z.; Qian, P. *J. Chem. Phys.* **2006**, *125*, 064311.

(55) Comptom, D. A.; Montero, S. M.; Murphy, W. F. *J. Phys. Chem.* **1980**, *84*, 3587–3591.

(56) Pitzer, K. S. *J. Chem. Phys.* **1944**, *12*, 310–314.

(57) Hwang, M.-J.; Stockfisch, T. P.; Hagler, A. T. *J. Am. Chem. Soc.* **1994**, *116*, 2515–2525.

(58) Durig, J. R.; Comptom, D. A. C. *J. Phys. Chem.* **1979**, *83*, 265–268.

(59) Lide, D. R., Jr.; Mann, D. E. *J. Chem. Phys.* **1958**, *29*, 914–920.

(60) Lowe, J. P. *Prog. Phys. Org. Chem.* **1968**, *6*, 1–80.

(61) Halgren, T. A.; Nachbar, R. B. *J. Comput. Chem.* **1996**, *17*, 587–615.

(62) Allinger, N. L.; Rahman, M.; Lii, J.-H. *J. Am. Chem. Soc.* **1990**, *112*, 8293–8307.

(63) Wang, Z. X.; Duan, Y. *J. Comput. Chem.* **2004**, *25*, 1699–1716.

(64) Hobza, P.; Kabeláč, M.; Šponer, J.; Mejzlík, P.; Vondrášek, J. *J. Comput. Chem.* **1997**, *18*, 1136–1150.

(65) Foloppe, N.; Mackerell, A. D. *J. Comput. Chem.* **2000**, *21*, 86–104.

(66) Nakagawa, S. *J. Comput. Chem.* **2007**, *28*, 1538–1550.

(67) Guan, Q. M.; Yang, Z. Z. *J. Theory Comput. Chem.* **2007**, *6*, 731–744.

(68) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. *J. Comput. Chem.* **2005**, *26*, 915–931.

CT9006647

JCTC Journal of Chemical Theory and Computation

Electric-Field-Assisted Electron Transfer in a Porphine−Quinone Complex: A Theoretical Study

Pekka J. Aittala,*,† Oana Cramariuc,‡ and Terttu I. Hukka*,§

Department of Chemistry and Bioengineering, Tampere University of Technology, P.O. Box 541, 33101 Tampere, Finland, Department of Physics, Tampere University of Technology, P.O. Box 692, 33101 Tampere, Finland, and IT Center for Science and Technology, Av. Radu Beller 25, Bucharest, Romania

Received July 7, 2009

Abstract: The effects of a static external electric field on the ground state electronic structure of a porphine−quinone (PQ) complex have been studied by using density functional theory (DFT). The energies of the excited states have been calculated with time-dependent density functional theory (TDDFT) and with the approximate coupled cluster singles and doubles (CC2) method. The geometries of porphine and quinone have been optimized with B3LYP. The influence of the external electric field on the PQ complex has been studied at six different intermolecular distances between 2.5 and 5.0 Å with the BH&HLYP functional. An external electric field clearly affects the orbitals localized mostly on quinone but not the orbitals localized on porphine. Additionally, the effect of the external field increases with the increasing intermolecular distance. The optical absorption spectrum of porphine obtained by using the BH&HLYP functional is consistent with the Gouterman model and with the spectrum previously calculated with CAM-B3LYP. The potential energy curves of the Q and B states and the lowest charge transfer (CT) states of the PQ complex calculated by using the BH&HLYP with TDDFT functional have also been compared with those obtained with the CC2 method. Both methods show that the lowest CT state is clearly above the Q states when no external field is applied. Therefore, when the Q states of a porphine−quinone system are excited, the conical intersection is not possible and cannot thus provide a path for electron transfer (ET). The calculations show that the Q and B states are affected by the field much less than the lowest CT state. Consequently, the calculations show that the CT state crosses the Q and B states at certain field strengths. Thus, it is possible that the external electric field triggers ET in porphine−quinone systems via conical intersection.

1. Introduction

Electron transfer (ET) plays a crucial role in the photosynthesis taking place in a reaction center. The initial step of photosynthesis is the photoexcitation of the light-harvesting chlorophyll, which forms a so-called special pair with a neighboring quinone, the whole system being embedded in protein. Once chlorophyll is photoexcited, an electron is transferred from chlorophyll to quinone. This initial ET in the special pair is a starting point of several subsequent electron transfer reactions in the reaction center.[1] The initial efficient ET step has attracted special attention over the years as scientists have tried to establish the mechanism of the ET reaction. Staab and co-workers[2−4] have synthesized relatively simple model systems for studying the first reaction step of photosynthesis composed of porphyrin derivatives linked covalently with cyclophane bridges to quinone for studying the photoinduced ET. They have found out that, when the porphyrin moiety of the zincporphyrin−quinone donor−acceptor dyads is photoexcited into the energetically

* Corresponding author tel.: +358(0)331153636; e-mail: pekka.aittala@tut.fi.

† Department of Chemistry and Bioengineering.

‡ Department of Physics.

§ IT Center for Science and Technology.

10.1021/ct9003417
Published on Web 02/05/2010

lowest excited state (the Q state), its fluorescence is quenched on a picosecond timescale. This implies efficient electron transfer from the porphyrin to the quinone, which has also turned out to be insensitive for the solvent environment.[3] On the contrary, in dyads consisting of a free-base porphyrin (porphine) and quinone, the fluorescence lifetime is strongly dependent on the polarity of the surrounding solvent. In polar solvents, the lifetime is on the same order of magnitude as in dyads consisting of zincporphyrin, but in nonpolar hexane the fluorescence lifetime is almost 10 ns, which is comparable to the lifetime of an unquenched porphine. Therefore, it has been concluded that there is no ET taking place from porphine to quinone in nonpolar solvents.[2]

Also, in several theoretical studies, ET from porphine derivatives to quinone has been investigated.[5−8] Worth and Cederbaum have studied the potential energy curves of the B, Q, and CT states of a zincporphyrin−quinone complex along the intermolecular distance by using the CIS method. They concluded that the CT states are in the same energy region with the B states localized on zincporphyrin. Additionally, the energies of the CT states depend more strongly on the molecular geometry than those of the local porphyrin states. Therefore, the CT states are likely to cross the B states, and these crossing points are part of a conical interaction seam proposed as a mechanism of formation of the CT state.[5] Worth and Cederbaum continued their work with Dreuw and Head-Gordon by studying a covalently linked zincporphyrin−quinone dyad by using time-dependent density functional theory (TDDFT) combined with ΔDFT/CIS. Their calculations showed that two large-scale motions, the "swinging-bridge" and the "twist motion" of the quinone moiety with respect to porphyrin, may cause excited state crossings between the locally excited Q states and the porphyrin-to-quinone CT state. Thus, these motions can trigger ultrafast ET from zincporphyrin to quinone via conical intersection, which is a region of coordinate space where two potential energy surfaces meet with a certain topology.[6]

Zheng et al. applied the INDO/S method to compute the electronic couplings via the two-state generalized Mulliken−Hush (GMH) approach of π-stacked poprhyrin−bridge−quinone systems. The phenyl linkers were found to dominate the mediation of the donor−acceptor coupling and the relatively weak exponential decay of the rate with distance aroused from the compression of the π-electron stack.[7] Olaso-González et al. studied the chlorin−quinone complex using CASPT2 and CASSCF methods. They concluded that ultrafast ET in a chlorin−quinone complex is possible only if the relative orientation of the donor and acceptor molecules allows some overlap of the LUMOs of the molecules. They agreed with Worth et al.[6] that large-scale motions must take place in the photosynthetic reaction centers to fulfill the observed ultrafast ET.[8]

The surrounding environment usually has a significant influence on the molecular properties. For example, an electric field induced by ambient molecules, for example, by zeolites and peptides, is reported to affect the molecular geometry, molecular orbital distribution, and dipole moments.[9] Furthermore, high electric fields (up to 10^9 V/m) induced by the large dipole moments of peptides are proposed to affect the electron transfer taking place between a donor and an acceptor embedded in between parallel peptide chains.[10] In some theoretical studies, the effects induced by an external electric field on molecular and electronic properties have been investigated. These studies have been focusing both on single molecules[9,11−14] and on donor−acceptor dyads.[15,16] However, cofacial dyads and complexes have not been much studied theoretically under the influence of an external electric field.[17]

Density functional theory (DFT) has been established as an efficient and reliable method for studying the ground state properties of molecules and solid systems. Furthermore, TDDFT has proven to be a reliable and facile method for calculating the properties of excited states.[18] However, the exchange-correlation functionals used currently in TDDFT calculations are reported to suffer from some well-documented shortcomings. In particular, TDDFT is known to underestimate the long-range CT excitations. This so-called CT failure of TDDFT is a consequence of the local nature of both the local density (LDA) and gradient corrected (GGA) approximations to the exact energy functional, which do not contain the derivative discontinuities required by the exact exchange-correlation potential.[18,19] Inclusion of a fraction of nonlocal Hartree−Fock exchange has been shown to improve results, as exemplified by the use of the "half and half" functional BH&HLYP (BHandHLYP) in the investigation of several excitations with CT character.[20−22]

In the current study, we have used DFT and TDDFT to investigate the effect of the electric field, which can be also induced by ambient molecules, on the excited state properties of a cofacial porphine−quinone[3,4] (PQ) donor−acceptor complex by simulating the surrounding effects with a static external electric field. The molecular structure of the studied complex is presented in Figure 1. We have used the hybrid BH&HLYP functional, which incorporates a high fraction (50%) of HF exchange. For example, in the case of a chlorin derivative bateriochlorophyll *b*, this functional has been reported to yield a rather good excitation spectrum.[23] However, as the BH&HLYP is not expected to completely remedy the CT failure of TDDFT, we have complemented our study by employing also the approximate coupled cluster singles and doubles (CC2) method. The SV(P) basis set, which is roughly identical to 6-31G*, has been used in all density functional calculations. The 6-31G* basis set has been previously successfully applied to calculations of the excited states of an isolated porphine[15] and a zincporphyrin−quinone dyad.[6] In CC2 calculations, we have used the TZVP basis set, which has been reported to be adequate for CC calculations of low-lying excited states not having Rydberg character.[24]

The aim of this study has been two-fold. First, we have studied the performance of the BH&HLYP functional in calculating the optical absorption spectrum of porphine because the current GGA functionals and hybrid functionals with a low fraction of Hartree−Fock (HF) exchange have been reported to yield qualitatively incorrect spectra for porphyrins and chlorophylls.[15] The second and the more important aim has been to study the effect of an electric field

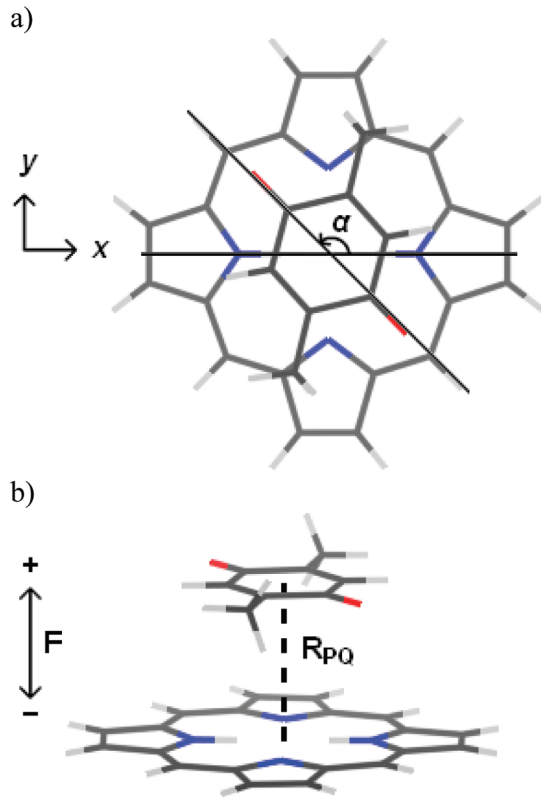


Figure 1. (a) Top view of a porphine−quinone complex. The rotation angle α is the angle between the axes drawn through the two oxygens of quinone and through the two "center" hydrogens of porphine. (b) Side view of the complex. R_{PQ} is the distance between the geometric center points of porphine and quinone. The external electric field (F) has been applied in the direction of a positive (+) or negative (−) z axis, that is, perpendicular to porphine and quinone planes. The direction of the external electrostatic field has been defined as a direction of the movement of a positive charge.

aligned perpendicularly to the porphine and quinone planes on the electronic properties of a PQ complex.

2. Computational Methods

The ground state geometries and the single-point energies were calculated by using DFT.[25,26] In geometry optimizations, Becke's three-parameter hybrid functional (B3LYP)[27−32] was applied, whereas in the single-point energy calculations, the "half-and-half" hybrid functional BH&HLYP[27−30,33] (BHandHLYP) was used. The BH&HLYP functional consists of 0.5(LDA + B88) + 0.5HF exchange and LYP correlation functionals. In all DFT calculations, the Karlsruhe split valence basis set with one set of polarization functions for all atoms except for hydrogen (SV(P))[34] was applied.

Vertical excitation energies were calculated with TDDFT[35−37] by using the BH&HLYP functional. Only the singlet states were considered. The computed transitions were transformed into simulated absorption spectra by applying a uniform Gaussian broadening with a standard deviation of 0.1 eV. The TDDFT-calculated excitation energies were compared to the energies calculated with the approximate coupled cluster singles and doubles (CC2)[38] method, employed in the RICC2 module[39−41] in Turbomole. In the CC2 calculations, the frozen-core approximation was employed for the 1s orbitals of the carbon and nitrogen atoms. In the excited state calculations, the SV(P) and TZVP[34] basis sets were applied, and in the CC2 calculations, an auxiliary basis set[42] of the same quality was used.

Table 1. Rotation Angles α (deg),[a] Optimized Intermolecular Distances R_{PQ} (Å), and the Relative Energies E (kJ/mol) of A−H Calculated at the DFT/B3LYP/SV(P) Level of Theory

conformer	α	R_{PQ}	E
A	0	4.01	1.6
B	15	4.05	2.2
C	45	3.89	1.8
D	60	3.93	2.6
E	90	3.92	3.2
F	110	3.92	2.6
G	120	3.85	1.3
H	135	3.82	0

[a] See Figure 1a for definition.

The order of the magnitude of the strength of the external electric field (10^9 V/m) considered in this study corresponds to the electric field observed in peptides, zeolites, and protein cavities. Additionally, we have studied fields of 1×10^9 V/m, 2×10^9 V/m, and 4×10^9 V/m. In the calculations including external electric field, the field (F) was applied in the direction of a positive (+) or negative (−) z axis, that is, perpendicularly to the porphine and quinone planes, see Figure 1 b. The direction of the external electrostatic field was defined as a direction of the movement of a positive charge. All calculations presented in this study were performed with the Turbomole 5.9−5.10 software packages.[43]

3. Results

3.1. Ground State Energies and Geometries. First, the geometries of quinone and porphine were optimized. Second, the two molecules were superimposed so that porphine was set on the xy plane by setting its geometric center point to the origin, and the z axis was directed through the geometric center point of quinone. Eight conformers (A−H) were made by rotating the quinone with respect to the z axis, that is, by changing the rotation angle α, see Figure 1a for the details. In all conformers, the intermolecular distance (R_{PQ}) was set to 2.5 Å. Thereafter, the conformers were optimized at the DFT/B3LYP/SV(P) level.

Rotation angles α, optimized intermolecular distances, and the relative energies are presented in Table 1. The absolute energies are given in Table S2 in the Supporting Information. In the lowest-energy conformer (H) presented in Figure 1b, the two methyl groups of quinone are almost on the top of the two free nitrogens of porphine, and the two hydrogens of quinone are almost on the top of the center hydrogens of porphine. The structural characteristics of conformer H are given in Table S1 in the Supporting Information. The energies of conformers A−H differ by 3.2 kJ/mol, at the most. A slight distortion from the lowest-energy geometry induces only a small change in energy, see the energies of G and H. The optimized intermolecular distance (R_{opt}) of the lowest-energy conformer H was found to be 3.82 Å at

Table 2. Complexation Energies (kJ/mol) of the PQ Complexes with Intermolecular Distances R_{PQ} (Å) of 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 Å Calculated under the Influence of an External Electric Field F of −4, −2, −1, 0, +1, +2, and +4 × 10^9 V/m at the BH&HLYP/SV(P)//B3LYP/SV(P) Level of Theory

	R_{PQ}					
F	2.5	3.0	3.5	4.0	4.5	5.0
−4	305.9	11.6	−7.2	−5.7	−1.6	0.2
−2	301.7	9.1	−9.2	−7.2	−2.7	−0.8
−1	299.3	7.6	−10.4	−8.1	−3.5	−1.4
0	296.5	5.8	−11.8	−9.1	−4.3	−2.0
−1	293.5	3.9	−13.3	−10.2	−5.1	−2.7
−2	290.3	1.7	−15.0	−11.5	−6.1	−3.4
−4	282.8	−3.3	−19.0	−14.5	−8.3	−5.1

the B3LYP/SV(P) level, and further calculations with BH&HLYP yielded 3.60 Å.

In the current study, one of the aims is to investigate the effect of the intermolecular distance R_{PQ} on the electronic properties of a porphine−quinone complex. Therefore, six conformers were made from the most stable PQ complex H in the following way. The optimized geometries of porphine and quinone were kept frozen, and the R_{PQ} was set to 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 Å. The complexes are denoted as $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$, respectively. Single-point energy calculations were carried out at the BH&HLYP/SV(P) level of theory for these structures.

Complexation energy, which indicates the stability of the complex, was calculated as the difference between the sum of the isolated porphine and quinone and the energy of the interacting PQ complex. Table 2 summarizes the complexation energies of the $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ complexes calculated under the influence of an external electric field of +4, +2, +1, 0, −1, −2, and −4 × 10^9 V/m. The orientation of the external electric field has been defined in section 2. Without the presence of the external field, the PQ complexes with $R_{PQ} \geq 3.5$ Å have negative complexation energies. The absolute value of the complexation energy of $PQ_{3.5}$, $PQ_{4.0}$, and $PQ_{4.5}$ exceeds the thermal energy at room temperature (∼2.48 kJ/mol), and porphine and quinone are therefore bound together in these complexes. The $PQ_{2.5}$, $PQ_{3.0}$, and $PQ_{5.0}$ would not exist without linkers because the thermal fluctuation would dissociate the complexes. It is also noteworthy that the single-point energy calculations predict the minimum of complexation energy to be 3.5 Å, that is, close to R_{opt} (3.6 Å). This implies that the geometries of porphine and quinone are not much affected by the complexation. The DFT does not, however, take into account the van der Waals interactions, and calculations at the CC2/TZVP level indicate much more stable complexes with complexation energies of +13.4, −123.9, −97.7, −60.4, −35.4, and −20.7 kJ/mol for $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$, respectively. Hence, the CC2 calculations predict a minimum in the ground state potential energy curve (PEC) at a shorter R_{PQ} than DFT, and the minimum is also much steeper than the rather flat PEC obtained with DFT (see also Figure 5).

The complexation energies increase when the external field increases, see Table 2. Additionally, the external field affects less when the intermolecular distance R_{PQ} increases. When the external electric field increases by 1 × 10^9 V/m from −4 × 10^9 V/m to +4 × 10^9 V/m, the complexation energy of $PQ_{2.5}$ increases by 2.1−3.6 kJ/mol. In $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$, an increase of the external field by 1 × 10^9 V/m increases the complexation energy by 1.3−2.5, 1.0−2.0, 0.8−1.5, 0.6−1.1, and 0.5−0.9 kJ/mol, respectively. Therefore, the complexation energies are affected more by the intermolecular distance between the porphine and quinone than by the external electric field.

3.2. Ground State Electronic Structures. *Without the External Electric Field.* The localizations and energies of the molecular orbitals (MOs) reveal the nature of the excited states and provide insight into the absorption spectra of the PQ complexes. Therefore, we will concentrate here on the orbitals that are involved in transitions giving rise to the Q, B, and the lowest CT bands of the PQ complexes, see section 3.3. Additionally, isoamplitude surfaces, orbital energies, as well as variation of the orbital energies in porphine−quinone complexes as a function of the external electric field of the HOMO−2, HOMO−3, HOMO−4, LUMO+3, and LUMO+4 orbitals is provided in the Supporting Information. In order to illustrate the differences in localizations of the MOs, isoamplitude surfaces of HOMO−1, HOMO, LUMO, LUMO+1, and LUMO+2 of $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, and $PQ_{4.5}$ are presented in Figure 2. Localization of the orbitals of $PQ_{5.0}$ is practically identical to that of the $PQ_{4.5}$ complex, and the isoamplitude surfaces of $PQ_{5.0}$ have been thus omitted. In complexes with short intermolecular distances ($PQ_{2.5}$ and $PQ_{3.0}$), the interaction of porphine and quinone increases the delocalization of the orbitals. Most of the orbitals of these complexes are delocalized over the whole complex and cannot be clearly related to the orbitals of either the isolated porphine or quinone. However, in $PQ_{3.5}$ the delocalization of the orbitals is decreased, and the orbitals of $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ are almost entirely localized on either porphine or quinone. In these complexes, the electronic structure is roughly the combination of the orbitals of the isolated porphine and quinone.

Comparison of the isoamplitude surfaces of the other complexes with those of $PQ_{3.5}$ reveals that LUMO and LUMO+2 cross, that is, change places, when the intermolecular distance increases to 4.0 Å, see Figures 2 and 3. These orbitals retain their reversed order when the intermolecular distance increases further.

The HOMO and HOMO−1 orbitals of the PQ complexes arise from the degenerate HOMO and HOMO−1 of porphine. The LUMO and LUMO+1 of the $PQ_{2.5}$, $PQ_{3.0}$, and $PQ_{3.5}$ complexes and LUMO+1 and LUMO+2 of the $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ complexes arise from the degenerate LUMO and LUMO+1 of the isolated porphine. The LUMO+2 of the $PQ_{2.5}$, $PQ_{3.0}$, and $PQ_{3.5}$ complexes and LUMO of the $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ complexes can be related to LUMO of the isolated quinone. The energies of the two highest occupied and three lowest unoccupied molecular orbitals of $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ are presented in Table 3. Additionally, the energies of the corresponding orbitals of the isolated porphine and quinone are shown. The energies of the orbitals localized on porphine in $PQ_{3.5}$ differ

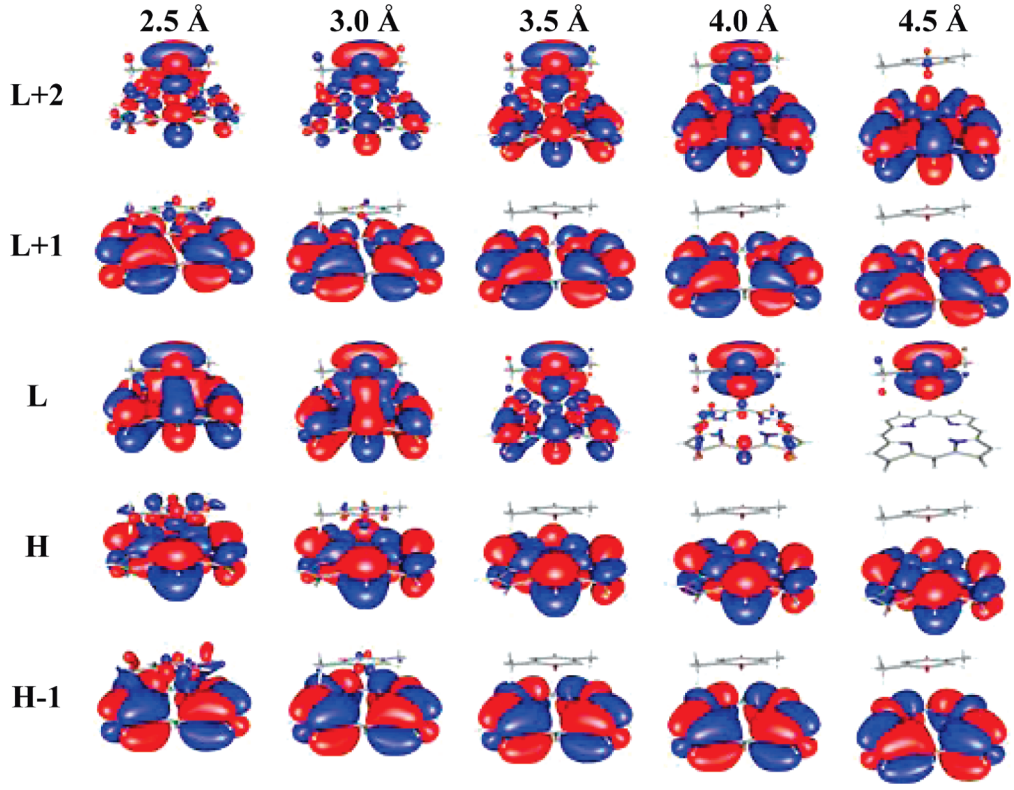


Figure 2. Some of the orbitals of the PQ complexes with intermolecular distances of 2.5, 3.0, 3.5, 4.0, and 4.5 Å calculated at the BH&HLYP/SV(P)//B3LYP/SV(P) level of theory without an external electric field. The isoamplitude surfaces of the orbitals presented are 10% of the maximum positive (red) and minimum negative (blue) amplitudes of the wave functions.

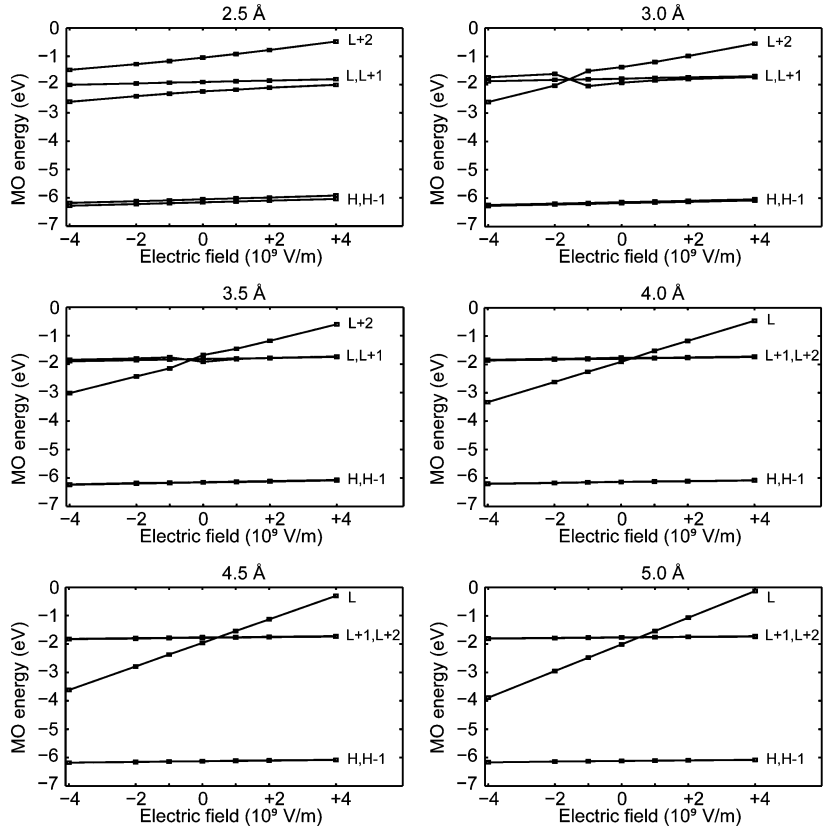


Figure 3. Variation of the orbital energies in porphine−quinone complexes with intermolecular distances of 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 Å as a function of the external electric field. See Figure 1b for a definition of the direction of the field. Orbitals are labeled according to the PQ complexes in the zero field. The energies are calculated at the BH&HLYP/SV(P)//B3LYP/SV(P) level of theory.

Table 3. Energies (eV) of Some Molecular Orbitals of the PQ Complexes with Intermolecular Distances R_{PQ} (Å) of 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 Å and the Energies of the Corresponding Orbitals of the Isolated Porphine (P) and Quinone (Q)[a]

	R_{PQ}							
orbital	2.5 Å	3.0 Å	3.5 Å	4.0 Å	4.5 Å	5.0 Å	P	Q
LUMO+2	−1.05	−1.38	−1.68	−1.76	−1.76	−1.76	0.32	1.84
LUMO+1	−1.91	−1.79	−1.82	−1.80	−1.78	−1.77	−1.72	0.79
LUMO	−2.24	−1.93	−1.92	−1.91	−1.96	−2.01	−1.74	−2.27
HOMO	−6.05	−6.14	−6.15	−6.14	−6.13	−6.12	−6.08	−9.00
HOMO−1	−6.16	−6.17	−6.16	−6.14	−6.13	−6.12	−6.08	−9.23

[a] The energies are calculated at the BH&HLYP/SV(P)//B3LYP/SV(P) level of theory.

only very little (<0.09 eV) from the energies of the corresponding orbitals of the isolated porphine. However, the energies of the orbitals localized entirely on quinone in the PQ complex are ∼0.5 eV smaller than the energies of the corresponding orbitals of the isolated quinone.

Changing of the intermolecular distance affects the energies of the MOs in which the isoamplitude surface is either delocalized to both porphine and quinone or localized entirely on quinone (see Figure 2 and Table 3), whereas the energies of the MOs localized on porphine remain practically constant. Except for $PQ_{2.5}$, the HOMO and HOMO−1 orbitals of the other complexes, localized on porphine, are degenerate, and the energies of these two orbitals do not depend on the intermolecular distance.

The LUMO is delocalized over the whole complex in $PQ_{2.5}$ and $PQ_{3.0}$ and mainly or entirely localized on quinone at 3.5 ≤ R_{PQ} ≤ 5.0 Å. The $PQ_{2.5}$ complex has the lowest LUMO energy. At 3.0 ≤ R_{PQ} ≤ 4.0 Å, the orbital energy is constant but decreases in $PQ_{4.5}$ and $PQ_{5.0}$ when the interaction between porphine and quinone decreases.

Regardless of the intermolecular distance, the LUMO+1 orbital is entirely localized on porphine, and the energy of the orbital is about the same. At 3.0 ≤ R_{PQ} ≤ 5.0 Å, the energy of the orbital changes by 0.08 eV at most.

In $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$, the LUMO+2 orbital is localized mainly on porphine. Thus, the energy of the orbital changes only a little between 4.0 and 5.0 Å. In $PQ_{3.5}$, $PQ_{3.0}$, and $PQ_{2.5}$, the decrease of the intermolecular distance leads to a stronger interaction between porphine and quinone, and the LUMO+2 becomes delocalized over the whole complex. The decrease of the intermolecular distance from 4.0 to 3.5 Å increases the orbital energy by 0.08 eV. When the intermolecular distance is decreased to 3.0 Å and further to 2.5 Å, the energy decreases by 0.30 and 0.33 eV, respectively.

The Effect of the External Electric Field. Variation of the energies of the two highest occupied and three lowest unoccupied MOs of the $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ complexes is presented as a function of the external electric field in Figure 3. Orbitals are labeled in the same way as in the case of the complexes in a zero field. Generally, the effect of the external electric field on the energies of the molecular orbitals delocalized over the whole complex or localized on quinone increases when the intermolecular distance increases.

In $PQ_{2.5}$, in which porphine and quinone interact strongly, the external electric field affects the orbitals only slightly. Therefore, neither the negative nor the positive electric field induces crossing, that is, no changes in the order of MOs. The orbital which is more strongly localized on quinone than on porphine, that is, LUMO+2, is affected the most. An increase of the positive electric field strength by 1×10^9 V/m increases the LUMO+2 energy by 0.1−0.16 eV. The negative electric field affects the energies of the LUMO slightly more than the positive field. The increase of the negative field strength by 1×10^9 V/m decreases the orbital energy by ∼0.1 eV, whereas the increase of the positive field increases the energy by ∼0.06 eV. Every increase of 1×10^9 V/m in the electric field strength increases the energy of the HOMO−1, HOMO, and LUMO+1 orbitals by 0.04 eV at the most.

Evidently, already in the $PQ_{3.0}$ complex, the orbitals localized mainly on porphine and mainly on quinone respond differently to the external electric field. The energy of the LUMO+1 increases by 0.03 eV at the most with an increase of the electric field by 1×10^9 V/m. The LUMO+2 crosses the LUMO when the external field is increased from -1×10^9 to -2×10^9 V/m, see Figure 3. This causes some nonlinear variation in the LUMO energy, but otherwise the energy of the orbital is affected by the external field in the same manner as the LUMO+1 localized on porphine by the external electric stimulation. On the contrary, the LUMO+2 that mostly localizes on quinone is affected much more. Every increase of 1×10^9 V/m in the strength of the electric field increases the energy of the LUMO+2 by ∼0.2 eV, but the crossing with LUMO between -1×10^9 and -2×10^9 V/m induces some exceptions to the linear behavior, see Figure 3. Energies of the degenerate HOMO and HOMO−1 are affected in the same way as the energy of the LUMO+1; that is, when the strength of the electric field increases by 1×10^9 V/m, the energy of the HOMO and HOMO−1 increase by 0.03 eV at the most.

In the complexes with longer intermolecular distances ($PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$), the external electric field changes the energies of the orbitals linearly. The energies of the orbitals localized on porphine are only slightly affected by the external field. With a few exceptions, the energies of the degenerate HOMO and HOMO−1 as well as the degenerate LUMO+1 and LUMO+2 (LUMO and LUMO+1 in $PQ_{3.5}$) increase only by 0.02 eV at the most, with a gradual increase of the strength of the external electric field. Therefore, the energies of the orbitals stay between −6.07 and −6.24 eV, −7.79 and −8.10 eV, and −1.72 and −1.90 eV, respectively (see Figure 3).

On the contrary, the energy of the LUMO (LUMO+2 in $PQ_{3.5}$) localized more strongly on quinone than on porphine is significantly affected by the external stimulation. Also, in this orbital, a linear dependence between the orbital energy and the external field is observed. The increase in the

Table 4. Energies *E* (eV) and Oscillator Strengths *f* of the Q and B Bands of Porphine Calculated with the TDDFT/BH&HLYP, TDDFT/CAM-B3LYP, and CC2 Methods[a]

	Q_x		Q_y		B_x		B_y	
	E	*f*	*E*	*f*	*E*	*f*	*E*	*f*
BH&HLYP/SV(P)	2.27	0.002	2.45	0.002	3.62	0.877	3.76	1.254
BH&HLYP/TZVP	2.25	0.003	2.43	0.003	3.59	0.939	3.69	1.247
CC2/SV(P)	2.30	0.001	2.70	0.003	3.56	0.982	3.65	1.200
CC2/TZVP	2.28		2.67		3.49		3.56	
CC2/SVP[b]	2.32		2.71		3.57		3.66	
CAM-B3LYP/6-31G*[c]	2.2		2.4		3.5		3.6	
experimental[d]	1.98	0.02	2.42	0.07	3.33	1.15	3.33	

[a] Experimental values are shown for comparison. [b] Obtained from ref 45. [c] Adopted from ref 15. [d] Obtained from ref 46.

intermolecular distance increases the effect of the external field on the orbital energy, and therefore the lines representing the change of the orbital energy become steeper along with the increasing intermolecular distance. Consequently, the longer the intermolecular distance, the smaller the external electric field that is able to cause the crossing of the orbital localized mostly on quinone with the orbitals localized on porphine.

Because the unoccupied orbitals of the PQ complexes localized on quinone and on porphine respond differently to the external electric field, it is expected that the electric field on the order of magnitude under consideration has a significant influence on the excited states of the PQ complexes. More specifically, the excited states mostly localized on quinone versus on porphine are expected to be influenced by the electric field more than the states localized only on porphine. Therefore, on the basis of the ground state electronic structure, it is expected that the external electric field affects also the electron transfer in the PQ complexes.

3.3. Electronic Absorption Spectra. *Porphine.* Over 30 years ago, Gouterman presented a four-orbital model[44] that explained the characteristic Q and B bands of porphines and chlorophylls. The model is widely accepted, but modern TDDFT calculations have shown that one has to go beyond the Gouterman model in order to explain all of the features of the porphine and chlorophyll spectra, for example, also the so-called N states. Among the traditional density functionals, BP86 and B3LYP have been reported as an improvement to the Gouterman model, but it has been recently shown[15] that only the spectra calculated with the computationally demanding CASPT2 method and the recent long-range corrected density functional, that is, CAM-B3LYP, are qualitatively consistent with experiments and support the Gouterman model. The N states, arising from the excitations from orbitals that are localized on two of the pyrrole rings only instead of the whole porphine, have been shown to have CT character. Hence, they respond to an external electric field applied along the porphine plane.[15] The N bands were proven to be significant in porphyrin spectroscopy, but the traditional density functionals underestimate the energies of the N bands clearly. Therefore, it has been concluded that in the TDDFT frame the CAM-B3LYP functional is mandatory for reliable theoretical investigations of the porphine absorption spectra.[15] It can be speculated whether the better performance of CAM-B3LYP compared to the traditional functionals is only due to the larger amount of the HF exchange or because of the range separation which improves the asymptotic behavior of the exchange potential of the functional. Therefore, we have studied the performance of the BH&HLYP functional, which does not contain the range separation, in calculating the optical absorption spectrum of porphine.

Table 4 summarizes the energies of the Q_x, Q_y, B_x, and B_y bands of porphine calculated by using the TDDFT/BH&HLYP and CC2 methods with the SV(P) and TZVP basis sets. In addition, the energies calculated using CC2/SVP[45] and TDDFT/CAM-B3LYP/6-31G*[15] are listed for comparison along with the experimental[46] values. The assignments of the bands were verified by applying a static external electric field of 2×10^9 V/m in the porphine plane, directed along the *x* and *y* axes, see Figure 1a. The states shown in Table 4 were not affected by the external field and can thus be confirmed as the Q_x (2.27 eV), Q_y (2.45 eV), B_x (3.62 eV), and B_y (3.76 eV) bands. In contrast, the states lying above the B bands responded to the external stimulation and are thus assigned to the N states. These non-Gouterman states contain mainly excitations from orbitals localized on two of the pyrrole rings only (from HOMO−2 and HOMO−3 to LUMO and LUMO+1).

In agreement with the Gouterman four-orbital model, TDDFT combined with BH&HLYP includes only the HOMO−1, HOMO, LUMO, and LUMO+1 orbitals in the transitions forming the first four excitation bands. Additionally, the energies of the Q and B bands calculated with BH&HLYP are almost identical to those calculated by using the CAM-B3LYP functional.[15] The BH&HLYP functional combined with the SV(P) basis set overestimates the experimental[46] energies of the Q and B bands by 0.43 eV at most, which is slightly less than the 0.5 eV reported in a study of the bacteriochlorophyll *b*.[23] The BH&HLYP/SV(P) yields the lowest two N states at 3.97 and 4.33 eV, that is, at slightly lower energies than CAM-B3LYP (∼4.3 and 4.5 eV). Compared to the results obtained with the SV(P) basis set, the use of the TZVP basis set with BH&HLYP decreases the energies of the Q states by 0.02 eV and those of the B_x and B_y states by 0.03 and 0.07 eV, respectively. Our findings confirm that, although the BH&HLYP functional overestimates the energies of the Q and B bands slightly, the spectrum is qualitatively consistent with the experiments and is in agreement with the results obtained by CAM-B3LYP and CASPT2. Therefore, we conclude that BH&HLYP is also suitable for investigating the properties of porphine derivatives.

The energies of the Q and B bands calculated by using the CC2 method with the SV(P) basis set are equal to the ones reported by Parac and Grimme[45] and do not differ much from the ones calculated with TDDFT/BH&HLYP/SV(P). Additionally, the two methods yield very similar oscillator strengths for these states. The CC2/SV(P) yields the Q_x and Q_y (2.30 and 2.70 eV) bands at slightly higher energies than TDDFT/BH&HLYP/SV(P), and thus the deviation from the experimental values increase. On the contrary, the CC2/SV(P)-calculated B_x and B_y energies (3.56 and 3.65 eV) are closer to the experimental values than the ones obtained by using TDDFT/BH&HLYP/SV(P). The use of the TZVP basis set instead of SV(P) in the CC2 calculations decreases the energies of the Q_x, Q_y, B_x, and B_y states by 0.02, 0.03, 0.06, and 0.09 eV, respectively.

Porphine−Quinone Complexes. The vertical singlet excitations of the isolated porphine and quinone were compared to the excitations of the $PQ_{2.5}$, $PQ_{3.0}$, $PQ_{3.5}$, $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ complexes. The simulated absorption spectrum obtained from the TDDFT/BH&HLYP/SV(P) calculations and the corresponding vertical excitations of $PQ_{2.5}$, $PQ_{3.0}$ and $PQ_{3.5}$ are presented in Figure 4a, b, and c, respectively. Except for an additional band, the spectrum of $PQ_{3.5}$ is in principle a superposition of the spectra of the isolated porphine and quinone (not shown). The simulated spectra of the $PQ_{4.0}$, $PQ_{4.5}$, and $PQ_{5.0}$ complexes are almost identical with that of $PQ_{3.5}$, and the assigning of the bands is straightforward. The increasing interaction between porphine and quinone changes the appearance of the spectra slightly in the case of $PQ_{3.0}$. Moreover, in the case of $PQ_{2.5}$, the absorption bands are clearly broader and the intensities of the B bands are smaller, whereas the CT band is more intense than in the spectra of the PQ complexes with longer intermolecular distances. However, in both cases, the identification of the Q, B, and CT states is clear. The use of the TZVP basis set instead of SV(P) affects the energies of the Q, B, and CT bands only very little, and thus the SV(P) basis set is used in the TDDFT calculations of the PQ complexes throughout the rest of this study.

The low-energy excitations at 2.25 eV (HOMO → LUMO+1 and HOMO−1 → LUMO) and 2.42 eV (HOMO−1 → LUMO+1 and HOMO → LUMO) in the spectrum of $PQ_{3.5}$ are identified as the Q_x and Q_y bands of porphine, respectively. The excitations with high oscillator strengths at 3.59 eV (HOMO−1 → LUMO+1 and HOMO → LUMO+2) and 3.67 eV (HOMO → LUMO+1 and HOMO−1 → LUMO+2) are identified as the Soret band (B_x and B_y, respectively) of porphine. The lowest states, which are localized on quinone, are found at 3.19, 3.47, and 4.00 eV. All excitations between 4.5 and 5 eV are localized porphine states. The band arising from two excitations having energies of 2.83 (HOMO → LUMO and HOMO → LUMO+2) and 2.90 eV (HOMO−1 → LUMO+2 and HOMO−1 → LUMO) is not present in the superposition of the simulated spectra of porphine and quinone and can thus be assigned to neither porphine nor quinone. The band is a consequence of the interaction between porphine and quinone and is identified as a porphine-to-quinone CT band.

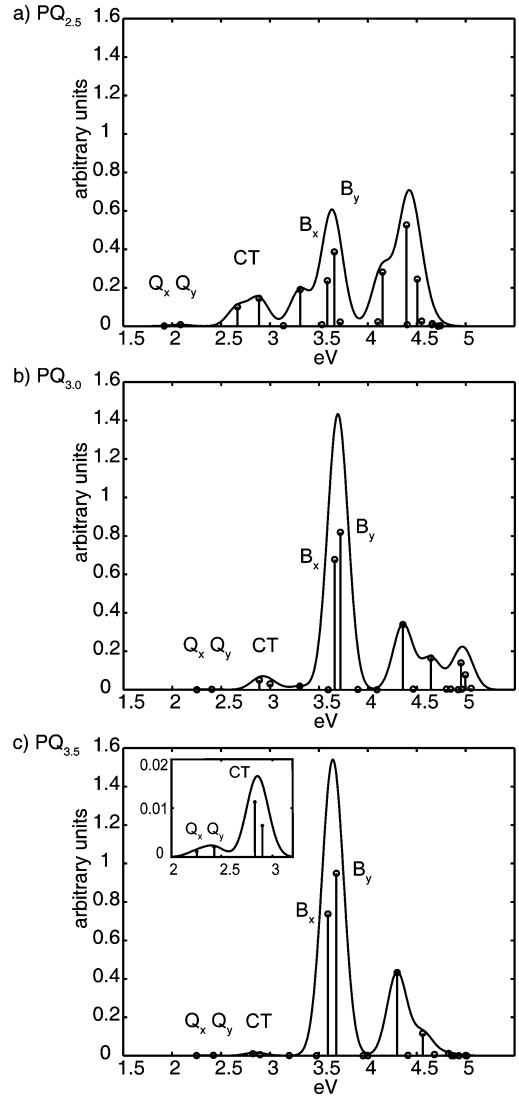


Figure 4. Simulated absorption spectra and the assignments of the Q, B, and CT bands of the PQ complexes with intermolecular distances of (a) 2.5 Å, (b) 3.0 Å, and (c) 3.5 Å calculated without an external electric field at the BH&HLYP/SV(P) level of theory.

3.4. Electron Transfer in Porphine−Quinone Systems. Recent theoretical studies of the photoinduced ET in zincporphyrin−quinone and in reduced chlorin−quinone systems have shown evidence that conical intersections (CIs) are involved in the underlying mechanism of this process.[5,6,8] The CIs are regions of coordinate space where two potential energy surfaces meet with a certain topology. Although between two states of the same multiplicity an infinite number of CI points form a crossing seam, the efficient transition from one state to the other occurs usually at the lowest CI. Thus the lowest minimum energy crossing point is of special interest and will constitute a key element in our investigation of the electric-field-mediated ET between porphine and quinone.[8]

In contrast to the zincporphyrin−quinone dyads, the fluorescence lifetime of the Q state of the porphine−quinone dyads[2] in nonpolar solvents is comparable to that of an unperturbed porphine, thus indicating that there is no ET from porphine to quinone. This would be evidenced by the

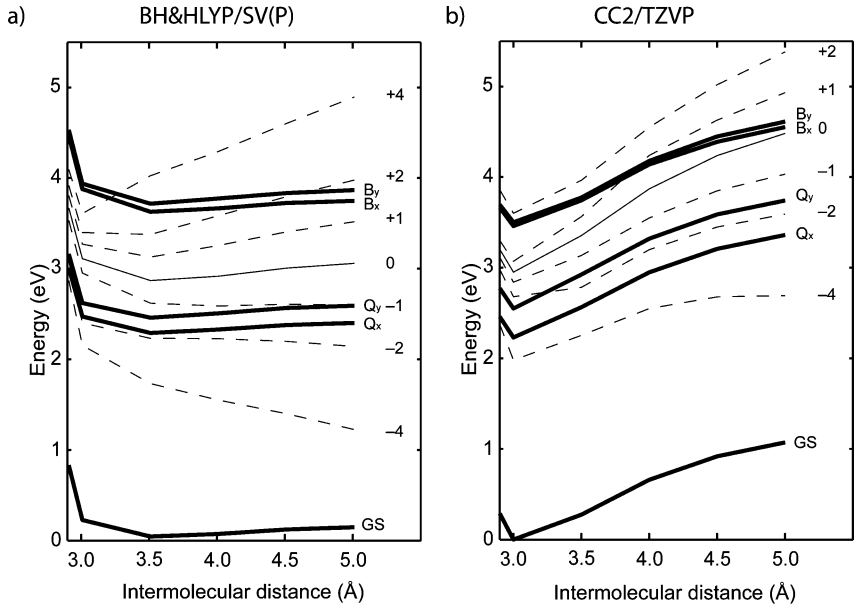


Figure 5. Potential energy curves (PECs) of the ground state (GS); the Q_x, Q_y, B_x, and B_y states (thick solid lines); and the energetically lowest CT state (thin solid line) as a function of the intermolecular distance of the PQ complex calculated without an external electric field. Thin dashed lines represent the PECs of the lowest CT state calculated under the influence of an external electric field of +4, +2, +1, −1, −2, and -4×10^9 V/m. The PECs of the Q and B states calculated in the presence of an external field were almost identical to the ones obtained in a zero field (thick solid lines) and have thus been omitted for clarity. The curves of the excited states are plotted relative to the GS curve calculated without the presence of an external field to allow easier comparison. The curves have been calculated at the (a) TDDFT/BH&HLYP/SV(P) and (b) CC2/TZVP levels of theory.

absence of a CI between the potential energy surfaces of the locally excited Q state and that of the CT state. Polar solvents decrease the fluorescence lifetime of porphine−quinone dyads significantly, which indicates that ET from porphine to quinone takes place in polar solvents[2] and a CI exists between the excited states. Thus, the question that arises is whether an external electric field can induce an ET between porphyrin and quinone in nonpolar solvents, leading to the crossing of the potential energy surfaces of the locally excited porphine states and the lowest CT state.

Without the External Electric Field. The PECs of the ground state; Q_x, Q_y, B_x, B_y states; and the energetically lowest CT state were calculated as a function of the intermolecular distance R_{PQ} of the PQ complex with the TDDFT/BH&HLYP and CC2 methods by applying the SV(P) and TZVP basis sets. The use of the TZVP basis set instead of SV(P) does not affect the energies of the PECs calculated with TDDFT/BH&HLYP. However, if the SV(P) basis set is used with CC2, the lowest CT state crosses the B_x state at $R_{PQ} = 5.0$ Å, but if TZVP is used instead the CT state remains below the B states at the whole R_{PQ} range studied. Hence, the SV(P) basis set can be used in the TDDFT calculations, but in the CC2 calculations the larger TZVP is needed. The PECs calculated at the (a) TDDFT/BH&HLYP/SV(P) and (b) CC2/TZVP levels of theory are presented in Figure 5. The PEC of the lowest CT state calculated with TDDFT/BH&HLYP does not cross either the Q or the B state (Figure 5a), thus indicating that no ET would occur when porphine is locally excited to any of these states. The CC2 calculations yield a similar picture for the Q state (Figure 5b), although the excited state energies are higher than the ones obtained with TDDFT/BH&HLYP. The CC2 method does not predict a crossing between the CT and B states in the studied R_{PQ} range, either. However, the CT state lies so close to the B states at 5.0 Å that the states could cross at slightly larger distances. Thus, ET from porphine to quinone could be possible, if the B states were excited, which could be a possible process, although this has not been investigated experimentally.

The Effect of the External Electric Field. The PECs of the energetically lowest CT state calculated under the influence of an external electric field of +4, +2, +1, 0, −1, −2, and -4×10^9 V/m are illustrated in Figure 5a and b as thin dashed lines. The curves of the excited states are plotted relative to the ground state curve calculated without the presence of an external field to allow an easier comparison.

Regardless of the direction and the strength, the external electric field does not practically affect the excitation energies of the Q and B states. Therefore, the potential energy curves of these states calculated in the presence of the external field are almost identical to the ones shown in Figure 5 as thick lines, which represent the PEC in the absence of the external field, and have been thus omitted for clarity.

Unlike the energies of the Q and B bands, the energy of the lowest CT band is clearly affected by the external stimulation. A positive orientation of the electric field shifts the energies of the CT state upward toward higher values, while a negative orientation shifts them down. Additionally, the stabilization or destabilization of the CT state is different at different intermolecular distances and depends also on the strength of the electric field. As a consequence, the PECs of CT can be influenced in such a way that they span a range of approximately 3 eV and have different slopes, thus intersecting the PEC of either the Q or the B states (see

Figure 5) at different intermolecular distances. This allows tuning of the electron transfer process as described below.

According to the TDDFT results, with the presence of an external field of $+2 \times 10^9$ V/m, the CT state crosses the B_x and B_y states at an R_{PQ} of ca. 4.3 and 4.7 Å, respectively. This means that under the influence of the positive field the CT state would closely resemble the ground state equilibrium structure, which would favor a fast electron transfer from porphine to quinone. A further increase of the electric field strength to $+4 \times 10^9$ V/m leads to the crossing of the CT and B states already in the inverted region at qn R_{PQ} of around 3.2−3.3 Å. Negative fields shift the PECs of the CT state further away from the B states and thus clearly hinders the ET if B states are excited. However, the CT state crosses the Q_y state already with the weakest studied negative field at $R_{PQ} \approx 5.0$ Å. The negative field could thus ease the forming of the CT state from the locally excited Q state. There is, however, a threshold after which the increasing of the negative field shifts the crossing of the Q and CT states to the inverted region. When stronger negative fields are applied, the CT crosses both Q states already at $2.5 < R_{PQ} <$ 3.0 Å, and the CT state becomes the lowest excited state. Although the general qualitative picture of the influence of the electric field on the CT formation is preserved also by the CC2 calculations, several quantitative conclusions are different.

The external electric field affects the energies of the Q and B states of the PQ complexes calculated by using CC2 (Figure 5b) by about the same amount as the energies calculated by using TDDFT/BH&HLYP. As discussed above, the CC2 calculations reveal that the lowest CT state lies so close to the B states at 5.0 Å that it is likely that the states cross at slightly large distances, when no external field is applied. This means that, if the B states are photoexcited in the zero field, the ET from porphine to quinone could be possible. According to the CC2 calculations, the smallest positive field shifts the crossing of the B states and the CT state to ∼3.8 Å. An increase of the field to $+2 \times 10^9$ V/m shifts the PEC of the CT state above the B states already at 2.5 Å, and no crossing occurs. In the R_{PQ} of 4.5 and 5.0 Å, the CT state calculated under the influence of $+4 \times 10^9$ V/m lies so much above the B states that the calculation of the CT state energies is not feasible. Thus, the corresponding PEC is totally omitted. However, this PEC is of less importance since it cannot cross the B states. The application of a negative electric field of -1×10^9 V/m decreases the CT state energy such that it lies clearly between the Q and B states, and no state crossings are observed at the intermolecular distances under consideration. When the negative field is increased to -2×10^9 V/m, the energy of the CT state decreases such that the CT state crosses the Q_y state at ca. 3.3 Å. Thus, CT formation from the Q states becomes possible. When the negative field is increased to -4×10^9 V/m, the CT state goes below the Q states already at 2.5 Å, and thus ET is not possible.

Comparison of the Methods. A comparison of our porphine−quinone calculations (TDDFT/BH&HLYP and CC2) with the CIS[5] and the combined TDDFT/BLYP and ΔDFT/CIS[6] calculations of the zincporphyrin−quinone system reveals differences between the two systems and between the performances of the methods. To begin with the locally excited porphine states, the BLYP yields the lowest locally excited quinone state in the zincporphyrin−quinone dyad clearly below the Q states. In our calculations, the lowest local quinone state is clearly above the Q states with an energy almost identical to the energy of the lowest excited state of the isolated quinone. This indicates, as expected, that the interaction is clearly stronger between quinone and zincporphyrin than between quinone and porphine. Moreover, the CIS calculations show that the interaction with quinone breaks the degeneracy of the B states of zincporphyrin, whereas in our CC2 calculations the perturbation of quinone actually increases the degeneracy of the B states of porphine.

Our CC2-calculated PEC of the lowest CT state of the porphine−quinone complex is rather similar to the one that has been calculated for zincporphyrin−quinone by using CIS.[5] Although the B states calculated with CC2 lie lower in energy than the states predicted by CIS, the CT state lies in both cases just below the B states at 5.0 Å, and it is likely that also in the case of the porphine−quinone complex the states would cross at about 5.5 Å, just like in CIS calculations of the zincporphyrin−quinone complex. A comparison of the CIS calculations with the ΔDFT/CIS calculations carried out by Dreuw and co-workers[6] reveals that the CIS method overestimates the energies of the CT states of the zincporphyrin−quinone system almost by 2 eVs. However, since porphine is a weaker electron donor than zincporphyrin, the CT states of the porphine−quinone system should lie higher in energy compared to the zincporphyrin−quinone system. Therefore, we expect that the CC2-calculated PEC of the CT state of the PQ complex is not much overestimated.

Comparison between the TDDFT/BH&HLYP and CC2 calculations indicates that both methods predict in principle a similar behavior for the PQ complex under the influence of the external electric field, and only the intermolecular distance, in which the locally excited porphine states (Q and B states) and CT states cross, changes. Considering the quite flat shape of the PEC of the CT state, it is clear that, despite the high fraction of the HF exchange (50%), TDDFT/BH&HLYP underestimates the CT energies at longer intermolecular distances. It could be that the CAM-B3LYP functional,[47] which has been reported to perform best with parameters that set the HF exchange to 65% in long distances, yields some improvement to the PECs of the lowest CT state. However, currently there is no study in which the performance of BH&HLYP and CAM-B3LYP is compared. The CAM-B3LYP functional has been reported to yield clearly better CT energies than B3LYP,[47] but also BH&HLYP has been reported to improve the CT calculations as compared to B3LYP.[22] On the basis of the porphine spectra calculated with BH&HLYP in this study and with CAM-B3LYP in another study,[15] we expect that the PECs of the Q and B states obtained with these two functionals would be very similar.

Regardless of the method (TDDFT/BH&HLYP or CC2), the calculations show that the CT states of the porphine−quinone complexes can be controlled by an external field

without perturbing the local porphine states. Therefore, an external electric field can be useful in controlling ET in porphine−quinone systems.

5. Conclusions

The influence of a static external electric field on the order of magnitude of 10^9 V/m, corresponding to the magnitude of an electric field induced by large dipole moments of peptides, on the ground state electronic structure and the singlet excited state energies of PQ complexes has been studied by using DFT, TDDFT, and the CC2. Six different intermolecular distances between 2.5 and 5.0 Å have been investigated.

An external electric field affects the energies of the orbitals localized mostly on quinone, whereas the orbitals localized entirely on porphine are hardly affected. Moreover, the effect of the external field on the orbital energies increases when the intermolecular distance increases.

In the current study, we have also shown that BH&HLYP yields a qualitatively correct porphine spectrum in which the N states lie clearly above the Q and B bands. Moreover, the calculated spectrum is almost identical to the one obtained previously by using CAM-B3LYP.

The potential energy curves of the Q and B states and the lowest CT state were calculated as a function of the intermolecular distance of the PQ complex both in the absence and in the presence of an external electric field. Both field directions, that is, from porphine to quinone and from quinone to porphine, were considered. Regardless of the direction or the strength, the external electric field affects the energies of the Q and B states only slightly. On the contrary, the energy of the lowest CT state depends both on the strength and the direction of the external field as well as on the intermolecular distance. Both methods (TDDFT/BH&HLYP and CC2) show that, depending on the strength and direction, the external electric field is able to either induce or hinder crossing of the locally excited porphine states (Q and B) and the lowest CT state. Moreover, crossing can be induced to occur at geometries close to those of the ground state, which would facilitate fast electron transfer. Thus, we conclude that the external electric field can be used to control ET in porphine−quinone systems.

Acknowledgment. Prof. H. Lemmetyinen, the head of the Laboratory of Chemistry at Tampere University of Technology, is acknowledged for offering the facilities for this research. The CSC−IT Center for Science Ltd., governed by the Finnish Ministry of Education, is acknowledged for providing the computing resources. Financing of this research by the Academy of Finland and the Romanian National University Research Council (RP7/6/30.04.2008) is greatly appreciated.

Supporting Information Available: Absolute energies of the complexes A−H and bond lengths and bond angles of the porphine−quinone complex H optimized at the B3LYP/SV(P) and BH&HLYP/SV(P) levels of theory are provided. In addition, isoamplitude surfaces; orbital energies; and a variation of the orbital energies of the HOMO−2, HOMO−3, HOMO−4, LUMO+3, and LUMO+4 orbitals of the $PQ_{2.5}$−$PQ_{5.0}$ complexes as a function of the external electric field are provided. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hoff, A. J.; Deisenhofer, J. *Phys. Rep.* **1997**, *287*, 1–247.

(2) Heitele, H.; Pöllinger, F.; Häberle, T.; Michel-Beyerle, M. E.; Staab, H. A. *J. Phys. Chem.* **1994**, *98*, 7402–7410.

(3) Häberle, T.; Hirsch, J.; Pöllinger, F.; Heitele, H.; Michel-Beyerle, M. E.; Anders, C.; Döhling, A.; Krieger, C.; Rückemann, A.; Staab, H. A. *J. Phys. Chem.* **1996**, *100*, 18269–18274.

(4) Staab, H. A.; Hauck, R.; Popp, B. *Eur. J. Org. Chem.* **1998**, 631–642.

(5) Worth, G. A.; Cederbaum, L. S. *Chem. Phys. Lett.* **2001**, *338*, 219–223.

(6) Dreuw, A.; Worth, G. A.; Cederbaum, L. S.; Head-Gordon, M. *J. Phys. Chem. B* **2004**, *108*, 19049–19055.

(7) Zheng, J.; Kang, Y. K.; Therien, M. J.; Beratan, D. N. *J. Am. Chem. Soc.* **2005**, *127*, 11303–11310.

(8) Olaso-González, G.; Merchán, M.; Serrano-Andrés, L. *J. Phys. Chem. B* **2006**, *110*, 24734–24739.

(9) Rai, D.; Joshi, H.; Kulkarni, A. D.; Gejji, S. P.; Pathak, R. K. *J. Phys. Chem. A* **2007**, *111*, 9111–9121.

(10) Fox, M. A.; Galoppini, E. *J. Am. Chem. Soc.* **1997**, *23*, 5277–5285.

(11) De Biase, P. M.; Doctorovich, F.; Murdiga, D. H.; Estrin, D. A. *Chem. Phys. Lett.* **2007**, *434*, 121–126.

(12) Choi, Y. C.; Kim, W. Y.; Park, K. -S.; Tarakeshwar, P.; Kim, K. S.; Kim, T. -S.; Lee, J. Y. *J. Chem. Phys.* **2005**, *122*, 094706−1−094706−6.

(13) Tsukamoto, S.; Nakayama, T.; Aono, M. *Chem. Phys.* **2007**, *342*, 135–140.

(14) Ye, Y.; Zhang, M.; Zhao, J. *THEOCHEM* **2007**, *822*, 12–20.

(15) Cai, Z.- L.; Crossley, M. J.; Reimers, J. R.; Kobayashi, R.; Amos, R. D. *J. Phys. Chem. B* **2006**, *110*, 15624–15632.

(16) Dahlblom, M. G.; Reimers, J. R. *Mol. Phys.* **2005**, *103*, 1057–1065.

(17) Fang, Y.; Gao, S.; Yang, X.; Shuai, Z.; Beljonne, D.; Brédas, J. L. *Synth. Met.* **2004**, *141*, 43–49.

(18) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.

(19) Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697–12699.

(20) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.

(21) Liao, M. -S.; Lu, Y.; Scheiner, S. *J. Comput. Chem.* **2003**, *24*, 623–631.

(22) Magyar, R. J.; Tretiak, S. *J. Chem. Theory Comput.* **2007**, *3*, 976–987.

(23) Sundholm, D. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4265–4271.

(24) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110-1–134110-25.

(25) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346–354.

(26) Von Arnim, M.; Ahlrichs, R. *J. Comput. Chem.* **1998**, *19*, 1746–1757.

(27) Dirac, P. A.M. *Proc. R. Soc. London A* **1929**, *123*, 714–733.

(28) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385–390.

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

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

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

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

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

(34) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

(35) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454–464.

(36) Bauernscmitt, R.; Häser, M.; Treutler, O.; Ahlrichs, R. *Chem. Phys. Lett.* **1997**, *264*, 573–578.

(37) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433–7447. Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2004**, *121*, 12772–12773.

(38) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

(39) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(40) Hättig, C.; Hellweg, A.; Köhn, A. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1159–1169.

(41) Hättig, C.; Köhn, A. *J. Chem. Phys.* **2002**, *117*, 6939–6951.

(42) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(43) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(44) Gouterman, M.; Wagniére, G. H.; Snyder, L. C. *J. Mol. Spectrosc.* **1963**, *11*, 108–127.

(45) Parac, M.; Grimme, S. *J. Phys. Chem. A* **2002**, *106*, 6844–6850.

(46) Edwards, L.; Dolphin, D. H.; Gouterman, M.; Adler, A. D. *J. Mol. Spectrosc.* **1971**, *38*, 16–32.

(47) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

CT9003417

JCTC Journal of Chemical Theory and Computation

Theoretical Study of Vibrationally Averaged Dipole Moments for the Ground and Excited C=O Stretching States of *trans*-Formic Acid

Leif O. Paulson,[†] Jakub Kaminský,[‡] David T. Anderson,[†] Petr Bouř,*[,‡] and Jan Kubelka*[,†]

Department of Chemistry, University of Wyoming, Laramie, Wyoming 82071, and Institute for Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo nám. 2, 16610, Prague, Czech Republic

Received September 12, 2009

Abstract: Recent experimental studies of *trans*-formic acid (FA) in solid para-hydrogen (pH_2) highlighted the importance of vibrationally averaged dipole moments for the interpretation of the high-resolution infrared (IR) spectra, in particular for the C=O stretch (ν_3) mode. In this report, dipole moments for the ν_3 ground ($v = 0$) and excited ($v = 1$, 2, 3, and 4) anharmonic vibrational states in *trans*-FA are investigated using two different approaches: a single mode approximation, where the vibrational states are obtained from the solution of the one-dimensional Schrödinger equation for the harmonic normal coordinate, and a limited vibrational configuration interaction (VCI) approximation. Density functional theory (B3LYP, BPW91) and correlated ab initio (MP2 and CCSD(T)) electronic methods were employed with a number of double- and triple-ζ and correlation consistent basis sets. Both single mode and VCI approaches show comparable agreement with experimental data, which is more dependent on the level of theory used. In particular, the BPW91/cc-pVDZ level appears to perform remarkably well. Effects of solvation of FA in solid state Ar and pH_2 matrices were simulated at the BPW91/cc-pVDZ level using a conductor-like polarized continuum model (CPCM). The Ar and pH_2 solid-state matrices cause quite a substantial increase in the FA dipole moments. Compared to gas-phase calculations, the CPCM model for pH_2 better reproduces the experimental FA spectral shifts caused by interaction with traces of ortho-hydrogen (oH_2) species in solid pH_2. The validity of the single mode approach is tested against the multidimensional VCI results, suggesting that the isolated (noninteracting) mode approximation is valid up to the third vibrationally excited state ($v = 3$). Finally, the contribution of the ground anharmonic vibrational states of the remaining modes to the resulting ν_3 single mode dipole moments is examined and discussed.

1. Introduction

Electrostatic forces are responsible for the structure of molecules and complexes, their spectroscopic detection, and intermolecular interactions.[1,2] Since an overwhelming majority of molecules are polar, electric dipole moments dominate in interactions with electric fields and form a basis for the most fundamental models of the condensed phases.[3] Solute−solvent electrostatic interactions can have profound effects on the properties of studied molecules, including their spectroscopic signatures,[4] which provide insight into the properties of both the studied solute species and the surrounding solvent.[5]

Formic acid (FA) is a prototypical, highly polar molecule, which has been subject to numerous vibrational spectroscopic investigations in the gas phase,[6−10] solution, and several low-

* To whom correspondence should be addressed. E-mail: bour@uochab.cas.cz (P.B.); jkubelka@uwyo.edu (J.K.).

[†] University of Wyoming.

[‡] Academy of Sciences of the Czech Republic.

10.1021/ct900608t
Published on Web 01/28/2010

temperature matrix isolation host materials.[11−18] In our recent study of *trans*-FA in low temperature para-H_2 (pH_2) matrices using high-resolution FTIR spectroscopy,[18] we observed multiple closely spaced peaks for the C═O stretching fundamental, commonly denoted as ν_3,[10] as well as for the $2\nu_3$ overtone band. We proposed that these multiplets arise from preferential clustering of the quadrupolar ortho-H_2 (oH_2) species, which is always present in ppm concentrations in the pH_2 matrix, to the FA dopant molecule. The electrostatic interaction between the oH_2 quadrupole moment and the dipole moments of FA in the ground and excited ν_3 vibrational states causes the splitting of the FA ν_3 absorption.

Quantitative explanation of these spectral features requires the value of the molecular dipole moment of FA in the ν_3 ground and vibrationally excited states. While for the ground and the first ν_3 excited state the dipole moments have been experimentally measured,[8] they have not been reported for higher excited states. Furthermore, no experimental dipole moment data are available for FA in any solid matrix environment. Since the solvent, or matrix, can have pronounced effects on the vibrational frequencies and intensities, especially those associated with polar bonds such as C═O,[19−24] it is likely that the permanent dipole moments will also be sensitive to the solvent or matrix environment. In this report, we present theoretical calculations of the vibrationally averaged dipole moments for the ground and excited states of the ν_3 vibrational mode of FA in the gas phase, as well as in an Ar and pH_2 matrix environment, treated by an implicit polarized continuum model.

Theoretical determination of the vibrationally averaged dipole moments necessitates accurate modeling of the anharmonic vibrational states. Unfortunately, no universal approach exists for obtaining vibrational properties of sizable molecules beyond the harmonic limit.[25] In this study, two different approaches are explored. The first is based on a one-dimensional (1D) anharmonic vibrational energy and dipole moment function calculated for the ν_3 normal mode of the FA. The normal mode coordinate is treated as independent from all other vibrational degrees of freedom. This approach is analogous to the simple local mode model,[26] which has proven very useful in the investigations of overtone vibrations involving predominantly X−H stretching,[27,28] including the O−H stretch of FA.[9,29,30] Since our treatment is based on the normal mode, we use the term "single mode method" to distinguish this approach from the conceptually different local mode theory of molecular vibrations.[31−33] The second approach is a multidimensional limited vibrational configuration interaction (VCI) methodology,[25,34] which includes all vibrational coordinates. A third approach, the multidimensional degeneration-corrected second-order perturbation theory,[25] was also tested with very similar results to those obtained by VCI, as observed previously for other vibrationally averaged properties.[35]

2. Computational Methods

2.1. Single Mode Model. In the single mode picture[26] the vibrational states (wave functions) are found as solutions to the 1D Schrödinger equation:

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dR^2} + V(R)\right]\psi(R) = E\psi(R) \quad (1)$$

where m is the reduced mass and $V(R)$ is the potential energy as a function of the ν_3 normal coordinate R. We use the symbol "R" for the non-mass-weighted normal coordinate, to distinguish it from the mass-weighted normal coordinates (Q) used in the next section. R and m were obtained from harmonic vibrational calculations using the Gaussian 98/03 quantum chemistry package.[36] Density functional theory (DFT, B3LYP, and BPW91 functionals) and correlated ab initio [MP2 and CCSD(T)] methods were employed along with a number of basis sets (see Results). The effects of the solid-state Ar and pH_2 matrices were simulated using the conductor polarized continuum model (CPCM)[37,38] with the dielectric constant $\varepsilon = 1.43$ for Ar and $\varepsilon = 1.294$ for pH_2.[39] Additional CPCM parameters (default in Gaussian 03) were the united atom (UA0) topological model for the solute radii (2.125, 1.75, and 1.85 Å radii for the CH, O and OH groups, respectively), the solvent radius of 1.875 Å, and the average tesserae area of 0.2 Å^2.

At each level of theory, the FA geometry was fully optimized, followed by a harmonic vibrational frequency calculation. The optimized geometries are listed and compared to available experimental data in Supporting Information, Table S1. Energies $V(R)$ and dipole moments $\mu(R)$ were computed for a series of 49 structures [25 structures for CCSD(T)/aug-cc-pVTZ level] generated at discrete steps along the ν_3 normal mode displacement R. The points were quadratically distributed from −0.35 to +1.0 Å with respect to the energy minimum to ensure adequate sampling of the potential near its maximum curvature. Gaussian was used for all calculations except CCSD(T), for which the population analysis is not implemented and the energy surfaces and dipole moments were calculated using ACESII.[40]

The Schrödinger equation (eq 2) for the resulting potential energy profile was solved numerically using the grid variational method[41,42] with MATLAB (Mathworks Inc. Mattick, MA) codes written in-house. The wave function was expanded as a linear combination of coordinate grid points $|r_i\rangle$:

$$|\psi(R)\rangle = \sum_{i=1}^{N_{grid}} \psi(r_i)|r_i\rangle \quad (2)$$

Substituting eq 2 into 3 and applying the standard variational principle with respect to $\psi(r_i)$ subject to the normalization constraint lead to the following system of linear equations

$$\sum_{j=1}^{N_{grid}} (H_{ij} - E\delta_{ij})\psi(r_j) = 0 \quad (3)$$

where

$$H_{ij} = \langle r_i| -\frac{\hbar^2}{2m}\frac{d^2}{dr^2} + V(r)|r_j\rangle \quad (4)$$

The grid consisted of 401 points from $r_{eq} - 0.3$ Å to $r_{eq} + 0.5$ Å. A fifth-order finite difference method was used for the second derivative:

$$\langle r_i|\frac{d^2}{dr^2}|r_j\rangle = \frac{1}{\Delta r^2}\Big[-\frac{5269}{1800}\delta_{ij} + \frac{5}{3}\delta_{i\pm1,j} - \frac{5}{21}\delta_{i\pm2,j} + \frac{5}{126}\delta_{i\pm3,j} - \frac{5}{1008}\delta_{i\pm4,j} + \frac{1}{3150}\delta_{i\pm5,j}\Big] \quad (5)$$

The potential energy is diagonal, i.e. $\langle r_i|V(r_i)|r_j\rangle = V(r_i)\delta_{ij}$; values of $V(r_i)$ at the individual grid points were interpolated from 40 single-point Gaussian energies (above). The five lowest eigenvalues and eigenvectors of the Hamiltonian matrix (eq 5), corresponding to the vibrational states $v = 0$, 1, 2, 3 and 4, were calculated using iterative sparse matrix methods as implemented in MATLAB. Vibrationally averaged dipole moments were obtained as

$$\boldsymbol{\mu}_v = \langle\psi_v(R)|\boldsymbol{\mu}(R)|\psi_v(R)\rangle = \sum_{i=1}^{N_{grid}} \psi_v(r_i)\,\boldsymbol{\mu}(r_i)\,\psi_v(r_i) \quad (6)$$

The dipole moments at the grid points $\boldsymbol{\mu}(r_i)$ were again interpolated from the values obtained from the quantum mechanical calculations.

2.2. Multidimensional Anharmonic Calculations. All vibrational degrees of freedom were considered in the Taylor expansion of the potential in the (mass weighted) normal mode coordinates Q_i up to the fourth order:

$$V(Q_1,...,Q_M) = \sum_{i=1}^{M}\frac{\omega_i^2}{2}Q_i^2 + \frac{1}{6}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}c_{ijk}Q_iQ_jQ_k + \frac{1}{24}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{l=1}^{M}d_{ijkl}Q_iQ_jQ_kQ_l + ... \quad (7)$$

where all cubic and semidiagonal normal mode quartic constants (i.e., with two and more identical indices, such as d_{ijkk}) were considered, obtainable by back and forth normal-mode numerical differentiation of harmonic force fields; ω_i are the harmonic frequencies and $M = 3 \times$ number of atoms $- 6$. The harmonic force fields were obtained from Gaussian[36] at four levels of theory: B3LYP/6-311++G(d,p), BPW91/cc-pVDZ, MP2/6-311++G(d,p), and CCSD(T)/6-311++G(d,p). As for the single mode method, Ar and pH_2 matrices were included at BPW91/cc-pVDZ level by CPCM solvent model, with the same parameters as detailed above. The program S4[43,44] was used for the anharmonic computations, enabling vibrational configuration interaction (VCI) within the harmonic oscillator basis functions. To limit the size of the VCI Hamiltonian, the harmonic basis was restricted to the ground and first five excited state wave functions. The effects of the size of the harmonic basis including up to seven excited states were tested for the BPW91/cc-pVDZ level calculations (Supporting Information, Table S3).

The dipole moment $\boldsymbol{\mu}_v$ was calculated from the VCI wave function ψ_v for each selected state v as a quantum average

$$\boldsymbol{\mu}_v = \langle\psi_v|\boldsymbol{\mu}|\psi_v\rangle \quad (8)$$

where the molecular dipole moment $\boldsymbol{\mu}$ was expanded as

$$\mu_\beta = \mu_{0,\beta} + \sum_i P_{\beta,i}Q_i + \frac{1}{2}\sum_{i,j}D_{\beta,ij}Q_iQ_j \quad (9)$$

where $\beta = \{x, y, z\}$ and **P** are the first and **D** the second normal mode dipole derivatives. The tensor **P** was obtained from the Cartesian dipole derivatives **Π** (atomic polar tensors) as

$$P_{\beta,i} = \sum_{\lambda,\alpha}\Pi_{\beta,\lambda\alpha}S_{\lambda\alpha,i} \quad (10)$$

where **S** is the normal mode−Cartesian transformation matrix. The second derivatives **D** were obtained in normal modes from **P**, using a two-step differentiation formula,

$$D_{\beta,ij} = \frac{P_{\beta,i}(Q_j + \Delta) - P_{\beta,i}(Q_j - \Delta)}{2\Delta} \quad (11)$$

with $\Delta = 0.0022$ au for the CCSD(T) computation; however, the first ($P_{\beta,i}$) and diagonal second ($D_{\beta,ij}$ with $i = j$) dipole derivatives were obtained from a two-step numerical differentiation of dipoles at CCD/6-311++G(d,p) level; the off-diagonal second dipole derivatives ($D_{\beta,ij}$ with $i \neq j$) were neglected in this case.

3. Results

3.1. Single Mode Approximation. A typical one-dimensional (1D) potential and dipole moment function for the ν_3 normal mode (calculated at BPW91/cc-pVDZ level) are shown in Figure 1a. Potential energy and dipole moment functions for additional levels of theory are shown in the Supporting Information (Figure S1). The corresponding solutions of the 1-D Schrödinger equation (eq 1) for $v = 0-4$ are shown in Figure 1b. From these solutions, the vibrational parameters, frequencies, and spectral intensities, as well as the vibrationally averaged dipole moments, were obtained as detailed in the Computational Methods.

The computed vibrationally averaged dipole moments, along with the equilibrium structure values and available experimental data, are shown in Table 1. To highlight the resulting trends, the vibrationally averaged dipole moments are also plotted in Figure 2. All calculations predict the dipole moment to increase with the vibrational excitation, in agreement with the available experimental data for the ground and the first excited states. In all cases, the ground vibrational state dipole moment is greater than that for the minimum structure. The DFT methods generally overestimate the dipole moments, in particular with the augmented correlation consistent basis sets. The exception is the ground-state dipole computed with BPW91/cc-pVDZ, which is slightly lower (by ~0.002 D) than the experimental value and in the best overall agreement. The first vibrationally excited state dipole moments are without exception computed too high and generally with larger error than the ground state ones. The closest to experiment is again BPW91/cc-pVDZ, yielding a ~0.012 D greater value. The post-HF methods uniformly predict lower dipole moments than DFT with the same basis sets, but similar trends with respect to the basis

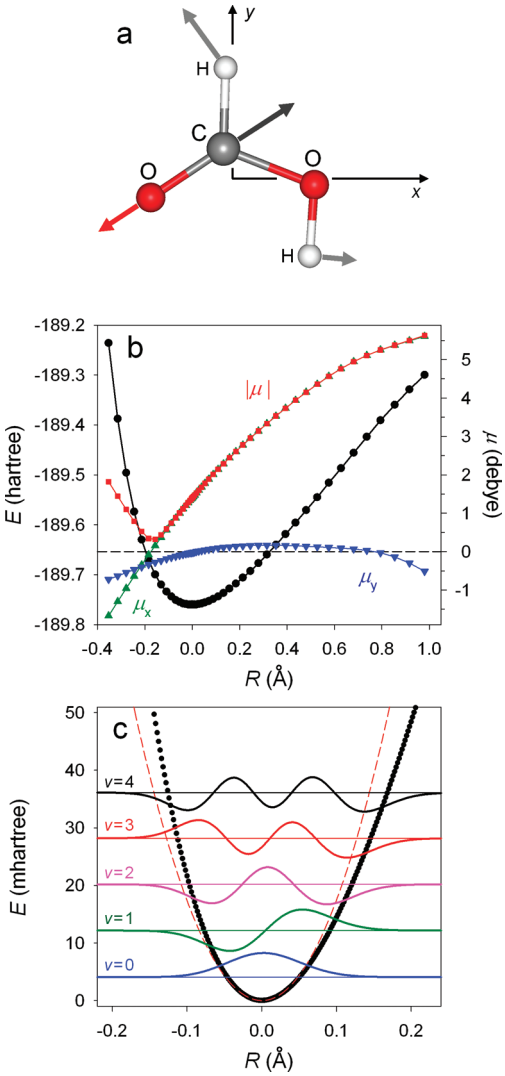


Figure 1. One-dimensional single mode representation of the ν_3 vibration in *trans*-FA. (a) The ν_3 normal mode of *trans*-FA. The molecule is oriented in the *x*−*y* plane, with axes parallel to the principal axes of the moment of inertia. The coordinate origin is at the center of nuclear charge. (b) Potential energy (black circles), the dipole moment components (*x*, green triangles; *y*, blue triangles), and magnitude (red squares) calculated at BPW91/cc-pVDZ level as functions of the ν_3 normal mode coordinate (*R*). (c) Solutions of the 1D Schrödinger equation for the potential energy function from part a shown as black circles. The dashed red line is the harmonic potential.

set size and type are observed. Namely, diffuse basis functions generally increase the computed dipole moments and augmented correlation consistent basis sets yield greater dipole values than 6-311++G(d,p). As a consequence, MP2 and CCSD(T) with the smaller basis sets severely underestimate the experimental dipole moments; however, the agreement improves with larger basis sets, in contrast to DFT. In particular, the CCSD(T)/aug-cc-pVTZ ground vibrational state dipole moment is in very good agreement with experiment (~0.005 D lower).

CPCM calculations for Ar and pH_2 matrices with BPW91/cc-pVDZ show a significant increase in the dipole moments with respect to the gas phase. The ground-state dipole moment in Ar is calculated to increase by ~0.16 D and in pH_2 matrix by ~0.11 D. A similar relative increase with respect to the gas phase is computed for all vibrationally excited states, the differences getting slightly smaller for higher v. These large matrix effects may seem somewhat surprising, given the rather subtle changes in FA geometry and vibrational frequencies (Supporting Information, Tables S1 and S6), but they are consistent with the increase in the spectral intensity (Supporting Information, Table S7).

As apparent from Figure 2, despite a wide variation in the dipole moments computed at various levels of theory, there is a systematic increase in the dipole moment with the vibrational state. It is therefore interesting to explore the changes in the computed excited state ($v = 1$, 2, 3, and 4) dipole moments with respect to the ground state ($v = 0$) as shown in Figure 3. The computed changes are systematically greater than the only available experimental reference (for $v = 1$), but more severe overestimation can be expected for the higher excited states. While the difference dipole moments are more consistent among all the methods, similar trends as those observed for the absolute values are still apparent: methods that gave higher absolute dipole moments also tend to give higher increments in the vibrationally excited states (Table 1, Figure 2). However, comparison of the dipole moment differences more clearly highlights the effects of basis sets. As evident from Figure 3, the same basis set yields relatively similar dipole moment changes irrespective of the method, and the greatest values are generally obtained with the largest (triple-ζ) augmented basis sets.

3.2. Multidimensional Anharmonic Calculations. The vibrationally averaged dipole moments obtained from the VCI anharmonic calculations for the ground and excited ν_3 states are summarized in Table 2. Generally, the VCI dipole moments are lower than those obtained in the single mode approximation. The difference is smallest for the ground vibrational state, about 0.05 D for all methods, but increases in the excited states. The $v = 0$ dipole moments are computed lower than those for the minimum energy structure, by ~0.02 D (DFT methods) and ~0.03 D (post-HF methods), in contrast to the single mode approximation, which systematically predicted greater vibrationally averaged dipole moments compared to the minimum energy structure. The most dramatic difference is predicted for the fourth ($v = 4$) excited state, where the VCI dipole moments are smaller, by approximately 0.25 D, than those obtained in the single mode approximation.

These differences are also reflected in trends with respect to the ν_3 vibrational state, as can be seen from Figure 4. Both DFT methods predict the dipole moment to increase up to $v = 3$, but the changes are significantly smaller than in the single mode approximation (Figures 2 and 3). The MP2 and CCSD(T) on the other hand yield a slightly smaller dipole moment for $v = 2$ than that for $v = 1$. While the gas-phase experimental dipole moment for $v = 2$ is not available, from the oH_2-induced frequency shifts in pH_2 matrix experiments[18] (also see below) it is evident that the dipole moment increases compared to the $v = 1$ (and $v = 0$) states. These qualitatively incorrect MP2 and CCSD(T) results may be explained by more anharmonic energy

Table 1. Vibrationally Averaged Dipole Moments (in D) for the ν_3 (C=O Stretch) Vibrational States in *trans*-FA in Single Mode Approximation

level	equilibrium	$v = 0$	$v = 1$	$v = 2$	$v = 3$	$v = 4$
B3LYP/6-31G(d,p)	1.4426	1.4601	1.4958	1.5318	1.5681	1.6047
B3LYP/6-31+G(d,p)	1.5315	1.5589	1.6132	1.6671	1.7207	1.7741
B3LYP/6-311++G(d,p)	1.4988	1.5245	1.5792	1.6337	1.6879	1.7419
B3LYP/cc-pVDZ	1.4243	1.4459	1.4894	1.5325	1.5753	1.6178
B3LYP/cc-pVTZ	1.5288	1.5755	1.6275	1.6793	1.7308	1.7821
B3LYP/aug-cc-pVDZ	1.5106	1.5346	1.5825	1.6302	1.6775	1.7245
B3LYP/aug-cc-pVTZ	1.5463	1.5710	1.6240	1.6768	1.7294	1.7817
BPW91/6-31G(d,p)	1.4181	1.4385	1.4797	1.5204	1.5609	1.6010
BPW91/6-31+G(d,p)	1.4867	1.5124	1.5640	1.6152	1.6661	1.7168
BPW91/6-311++G(d,p)	1.4592	1.4854	1.5383	1.5908	1.6430	1.6950
BPW91/cc-pVDZ	1.4026	1.4231	1.4645	1.5055	1.5461	1.5864
BPW91/cc-pVTZ	1.4911	1.5273	1.5771	1.6266	1.6759	1.7249
BPW91/aug-cc-pVDZ	1.4728	1.4957	1.5418	1.5875	1.6328	1.6779
BPW91/aug-cc-pVTZ	1.4984	1.5235	1.5742	1.6244	1.6744	1.7241
MP2/6-311++G(d,p)	1.3222	1.3440	1.3888	1.4329	1.4765	1.5197
MP2/cc-pVDZ	1.2763	1.2928	1.3266	1.3598	1.3923	1.4243
MP2/cc-pVTZ	1.3866	1.4059	1.4447	1.4828	1.5204	1.5574
MP2/aug-cc-pVDZ	1.4215	1.4426	1.4859	1.5286	1.5709	1.6129
MP2/aug-cc-pVTZ	1.4329	1.4557	1.4997	1.5433	1.5864	1.6291
CCSD(T)/6-311++G(d,p)	1.3307	1.3508	1.3981	1.4453	1.4921	1.5385
CCSD(T)/aug-cc-pVDZ	1.4281	1.4474	1.4938	1.5397	1.5854	1.6301
CCSD(T)/aug-cc-pVTZ	1.4041	1.4201	1.4693	1.5157	1.5632	1.6104
BPW91/cc-pVDZ/CPCM(Ar)	1.5583	1.5802	1.6256	1.6708	1.7156	1.7600
BPW91/cc-pVDZ/CPCM(pH_2)	1.5160	1.5348	1.5693	1.6046	1.6404	1.6766
experiment (gas phase)[a]	–	1.4253	1.4512	–	–	–

[a] Reference 8.

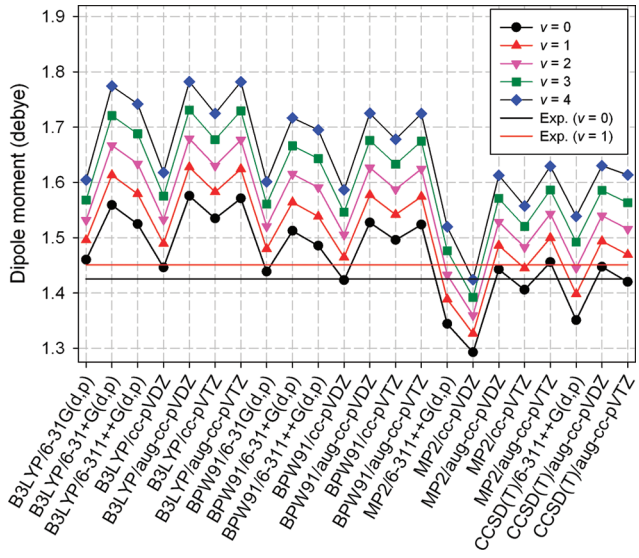


Figure 2. Dipole moments calculated in single mode approximation at various levels of theory for ground ($v = 0$, black circles) and excited ($v = 1$, red triangles; $v = 2$, pink triangles; $v = 3$, green squares; and $v = 4$, blue diamonds) ν_3 vibrational states. Experimental values for $v = 0$ (black) and $v = 1$ (red) are shown as solid lines.

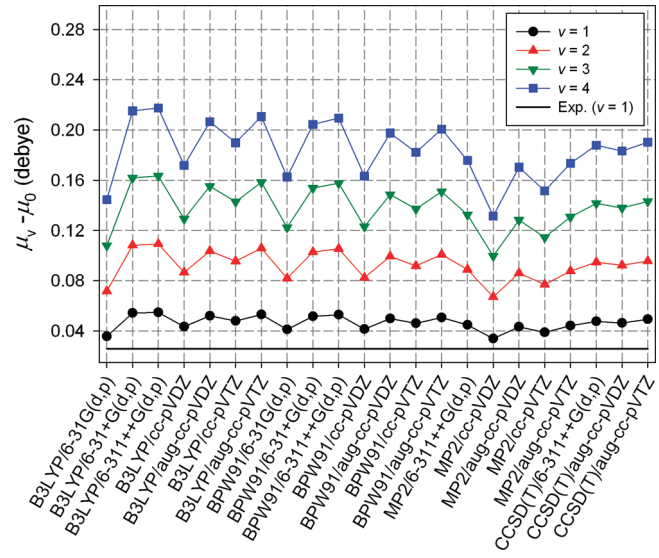


Figure 3. Difference dipole moments (with respect to $v = 0$) calculated in the single mode approximation at various levels of theory for excited ν_3 vibrational states ($v = 1$, black circles; $v = 2$, red triangles; $v = 3$, green triangles; and $v = 4$, blue squares) with respect to the ground state. The experimental value (for $v = 1$) is shown as a solid black line.

surfaces resulting from the wave function methods, compared to the DFT.[24,45] The limited VCI based on the post-HF calculations may therefore not be adequate even for the vibrational quantum numbers as low as $v = 2$. From $v = 2$ to $v = 3$ all methods compute a significant increase, followed by a rather dramatic decrease of the dipole moment from $v = 3$ to $v = 4$ vibrational states. These nonuniform trends, contrasting the fairly systematic single mode results, are particularly apparent from the difference dipole moment values (with respect to $v = 0$) in Figure 4b.

In comparison with experiment, the B3LYP/6-311++G(d,p) calculated dipole moments are too large, while those obtained with the other methods, in particular the post-HF, are too small (Figure 4a). The BPW91/cc-pVDZ values fall in between and, while lower than experiment, are again in the closest agreement. These results are consistent with the single mode approximation at the same levels of theory. Comparing the relative values of the first excited vibrational state ($v = 1$) dipole moments with respect to $v = 0$ (Figure 4b), the B3LYP/6-311++G(d,p) calculation almost exactly repro-

duces the experimental difference. The BPW91/cc-pVDZ, MP2/6-311++G(d,p), and CCSD(T)/6-311++G(d,p) all underestimate the dipole moment increase. This is again in contrast with the single mode results, where all the calculations systematically overestimated the experimental differences.

In the simulated solid matrices, an increase of ~0.17 D in Ar and ~0.12 D in pH_2 with respect to the gas phase is predicted for all vibrational states. These changes are nearly identical to the single mode results, only slightly larger, and again approximately correspond to the changes in the equilibrium structure dipole moments due to the reaction field of the matrix. Unlike the single mode calculations, however, the VCI differences between the gas and matrix phases show a slight increase, rather than decrease, with the vibrational quantum number.

3.3. Effects of Residual oH_2 Clustering in pH_2 Matrices. Finally, we return to the original motivation for this computational study: modeling the IR spectral frequency shifts due to the clustering of quadrupolar oH_2 to the FA in pH_2 matrix.[18] The interaction between the quadrupolar oH_2 and FA can be expressed as

$$V_{dq}(R,\theta_1,\theta_2,\varphi) = \frac{3\mu_1\Theta_2}{2R^4}[\cos\theta_1(3\cos^2\theta_2 - 1) + 2\sin\theta_1\sin\theta_2\cos\theta_2\cos\varphi] \quad (12)$$

where μ_1 is the dipole moment of molecule "1" (FA), Θ_2 the quadrupole moment of molecule "2" (oH_2, 0.194116 au), the angles are assumed to correspond to the minimum energy configuration ($\theta_1 = \theta_2 = \varphi = 0$), and $R = 3.79$ Å, the nearest neighbor spacing of the pH_2 crystal. Note that the minimum energy configuration refers to the orientation of the electric moments rather than a particular orientation of the oH_2 and FA molecules: the oH_2 quadrupole moment arises from the $J = 1$ state and therefore is inherently averaged over the $J = 1$ rotational wave function.

In the original paper,[18] the single mode BPW91/cc-pVDZ dipole moments computed in the gas phase were used to estimate the oH_2-induced frequency shifts. In the present study, the matrix effects were included using CPCM solvent model at a BPW91/cc-pVDZ level, which is computationally inexpensive and, as shown above, yields perhaps the best overall agreement with the gas phase as well as Ar and pH_2 matrix experimental data. Both single mode and VCI anharmonic calculations were performed, which allow us to reexamine the earlier results. Comparison with the experi-

Table 2. VCI Dipole Moments (in D) for the ν_3 (C=O stretch) Vibrational States in *trans*-FA

level	equilibrium	$v = 0$	$v = 1$	$v = 2$	$v = 3$	$v = 4$
B3LYP/6-311++G(d,p)	1.4988	1.4791	1.5044	1.5199	1.5488	1.4912
BPW91/cc-pVDZ	1.4026	1.3779	1.3884	1.4059	1.4158	1.3623
MP2/6-311++G(d,p)	1.3222	1.2893	1.2952	1.2915	1.3120	1.2688
CCSD(T)/6-311++G(d,p)[a]	1.3534	1.3270	1.3588	1.3422	1.3817	1.3139
BPW91/cc-pVDZ/CPCM(Ar)	1.5583	1.5422	1.5579	1.5716	1.5898	1.5306
BPW91/cc-pVDZ/CPCM(pH_2)	1.5160	1.4945	1.5103	1.5244	1.5408	1.4776
experiment (gas phase)[b]	–	1.4253	1.4512	–	–	–

[a] Molecular dipole moment calculated at the CCD/6-311++G(d,p) level. [b] Reference 8.

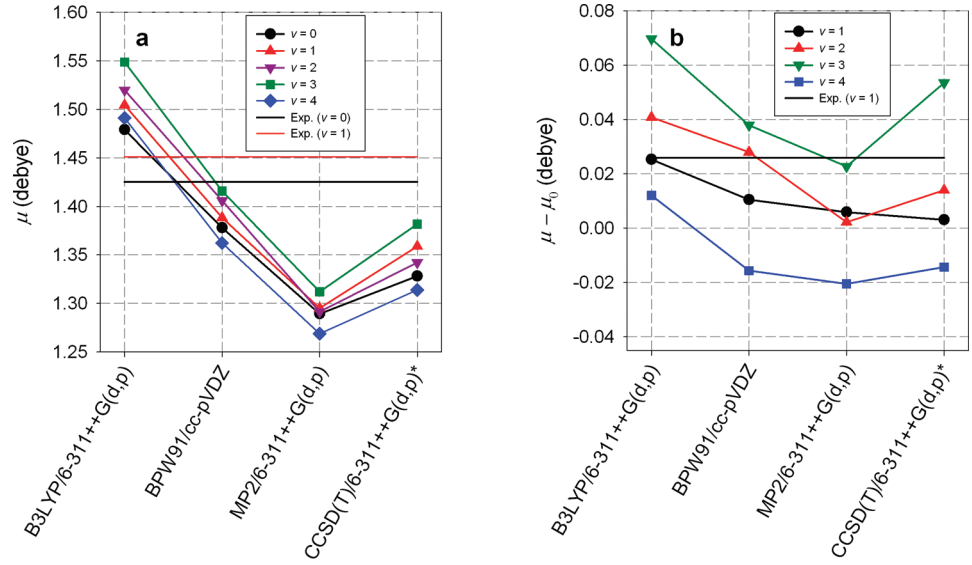


Figure 4. Dipole moments for ν_3 vibrational states of *trans*-FA calculated by VCI at various levels of theory. (a) Absolute dipole moments for the ground ($v = 0$, black circles) and excited ($v = 1$, red triangles; $v = 2$, pink triangles; $v = 3$, green squares; and $v = 4$, blue diamonds) ν_3 vibrational states calculated in single mode approximation at various levels of theory. Experimental values for $v = 0$ (black) and $v = 1$ (red) are shown as solid lines. (b) Difference dipole moments in excited ν_3 vibrational states ($v = 1$, black circles; $v = 2$, red triangles; $v = 3$, green triangles; and $v = 4$, blue squares) with respect to the ground state. Experimental value ($v = 1$) is shown as a solid black line.

Table 3. Interaction Energies (V_{dq}) and Frequency Shifts (Δ) Due to Quadrupole-Dipole Interaction between oH_2 and *trans*-FA in Ground and Excited ν_3 (C=O Stretch) Vibrational States

	gas phase				CPCM pH_2				
	single mode		VCI		single mode		VCI		
state	V_{dq} (cm^{-1})	Δ (cm^{-1})	V_{dq} (cm^{-1})	Δ (cm^{-1})	V_{dq} (cm^{-1})	Δ (cm^{-1})	V_{dq} (cm^{-1})	Δ (cm^{-1})	experiment:[a] Δ_{exp} (cm^{-1})
$v = 0$	27.28		26.41		29.42		28.86		
$v = 1$	28.07	−0.79	26.61	−0.19	30.08	−0.66	28.95	−0.30	−0.32
$v = 2$	28.86	−1.58	26.95	−0.54	30.79	−1.37	29.22	−0.57	−0.66

[a] Reference 18.

mentally observed frequency shifts due to oH_2 clustering also provides an additional, albeit indirect, experimental check of the computed dipole moments.

Interaction potentials for $v = 0$, 1, and 2 ν_3 vibrational states (eq 12) and the estimated oH_2-induced frequency shifts are summarized in Table 3. For comparison, the corresponding gas-phase results and experimentally observed frequency shifts are also included. The gas-phase dipole moments calculated in the single mode approximation yield frequency shifts that are considerably higher, while the VCI somewhat lower, than the experiment. This reflects the differences between the excited and ground vibrational state dipole moments computed by the single mode and VCI methods (Figures 3 and 4b). With the pH_2 matrix, approximated by the CPCM model, the single mode frequency shifts become smaller, while the VCI ones increase. Both result in better agreement with the experimental values, despite the qualitatively opposite effect of the matrix. This is a consequence of the subtle differences in the influence of the pH_2 reaction field on the dipole moments in the ground and vibrationally excited states: the VCI with the CPCM gives slightly steeper dipole moment dependence on the vibrational excitation than the corresponding gas-phase calculation, while the opposite is true for the single mode treatment. Therefore, it appears that CPCM indeed improves the experimental frequency shift prediction; however, given the very small magnitude of these changes and inaccuracies of the computational methods, these results have to be regarded with caution.

4. Discussion

The vibrationally averaged dipole moments obtained using two different approximations, the single mode and VCI anharmonic calculations, reveal some systematic differences. For example, the dipole moments from the VCI calculations are consistently lower than the single mode results. The single mode treatment also uniformly predicts a greater increase of the dipole moment with vibrational excitation than VCI. Comparison of the anharmonic frequencies and IR intensities, presented in the Supporting Information, also shows that the VCI yields systematically lower fundamental and first overtone ν_3 frequencies and infrared intensities (Supporting Information, Tables S6−S9). On the other hand, for the higher overtones the VCI predict higher vibrational frequencies, and a reversal in the relative intensities is also observed for the third ν_3 overtone. The overall agreement with the available experimental data, including frequencies and IR absorption intensities,[6−18] is, however, comparable for both the single mode and VCI approaches and is more significantly dependent on the particular level of theory used in the calculation. The VCI calculations tend to agree better with the experimental frequencies and intensities (Supporting Information), especially for the fundamental transitions. The single mode calculations, by contrast, give closer agreement with the experimental dipole moments.

The fundamental assumption of the single mode method is that the vibrational motion follows the normal mode coordinate, which is equivalent to the neglect of interactions with the other normal modes. Although the ν_3 is not expected to strongly interact with the other modes, since it is localized and energetically well separated from other transitions, it is likely that the single mode picture breaks down for higher excited vibrational states. To estimate the validity of the local mode assumption, we examine the dependence of the molecular dipole moment on the vibrational quantum number v. For an isolated mode the dipole moment increases linearly with the vibrational state.[46] In fact, this linearity has been used as evidence for the local character of the OH stretching in HOCl[46] and H_2O.[47,48] Therefore, it is not surprising that at all levels of theory the single mode treatment yields a near-perfect linear fit of the vibrationally averaged dipole moment as a function of the vibrational state (v) with correlation coefficients better than 0.9999. Linear extrapolation from the calculated dipole moments for the ground ($v = 0$) and first excited ($v = 1$) states to $v = 2$, 3, and 4 predicts the actual computed dipole moments to better than 0.0009 D for $v = 2$, 0.0022 D for $v = 3$, and 0.004 D for $v = 4$. Therefore, if the single mode approximation were valid, this near-perfect linear relationship would allow a straightforward estimation of the true dipole moments based on the known experimental values for the ground and first excited ν_3 vibrational states of FA (Table 1). With the most conservative estimation of the error, these values would be 1.4771 ± 0.0009 D for $v = 2$, 1.503 ± 0.002 D for $v = 3$, and 1.529 ± 0.004 D for $v = 4$.

Examination of the linearity of this relationship predicted by the VCI anharmonic calculations provides an independent test for the "locality" of the C=O stretching mode. Unfortunately, the dipole moment dependence on v in the VCI calculations varies widely, depending on the level of theory used. The DFT methods yield a very good linear relationship for $v = 0$, 1, 2, and 3 with the correlation coefficient of 0.994 for both B3LYP/6-311++G(d,p) and BPW91/cc-pVDZ. On the other hand, for the post-HF levels, there is hardly any linear trend. This qualitative inconsistency illustrates that the VCI method is also subject to inherent approximations and errors. The truncated Taylor expansions

Table 4. Dipole Moments (in D) for the ν_3 (C═O Stretch) Vibrational States in *trans*-FA in Single Mode Approximation, with the Correction for the ZPE States of the Remaining Modes and from VCI Anharmonic Theory at the BPW91/cc-pVDZ Level

method	$v = 0$	$v = 1$	$v = 2$	$v = 3$	$v = 4$
single mode	1.4231	1.4645	1.5055	1.5461	1.5864
single mode ZPE corrected	1.3935	1.4349	1.4759	1.5165	1.5568
anharmonic VCI	1.3779	1.3884	1.4059	1.4158	1.3623
experiment (gas phase)[a]	1.4253	1.4512	–	–	–

[a] Reference 8.

of the potential energy (eq 7) and dipole moment (eq 9) are expected to introduce errors, especially for higher excited vibrational states, which are more sensitive to the energy and dipole surfaces further away from the energy minimum. Moreover, the limited harmonic oscillator basis (to states where the sum of quantum numbers for all modes is less of equal to five) will also likely affect the results for the higher vibrationally excited state properties.

We have tested the convergence of the dipole moment values for $v = 0$ to $v = 4$ vibrational states, obtained with BPW91/cc-pVDZ parameters, with the size of the VCI harmonic basis, which included from four up to seven excited state wave functions (Supporting Information, Table S3). While the dipole moments in all vibrational states are somewhat dependent on the harmonic basis size, up to $v = 3$ the computed values are essentially stable. The $v = 4$ dipole moment, as might be expected, is the most sensitive to inclusion of additional basis functions. The qualitative trend of the decrease in the dipole moment from $v = 3$ to $v = 4$, however, is not affected. Furthermore, the larger the harmonic basis, the more significantly the $v = 4$ vibrational state becomes mixed with other harmonic modes. For example, with the largest harmonic basis tested, including up to seven excited harmonic functions, the contribution of the ν_3 normal mode to the VCI anharmonic wave function is only 27%. As a consequence, the $v = 4$ vibrational state can no longer be considered a pure ν_3 mode, which is consistent with the breakdown of the linear dependence of the dipole moment on the vibrational quantum number.

All combined, it is apparent that the fourth ($v = 4$) excited vibrational state cannot likely be treated as an isolated mode and the single mode approximation cannot be expected to yield reliable dipole moment values. On the other hand, as suggested by the DFT VCI results, up to $v = 3$ the approximation may hold, in which case the above prediction for the gas-phase $v = 2$ and $v = 3$ dipole moments in the neighborhood of 1.48 and 1.50 D, respectively, should be valid. The experimental study of the dipole moments of OH stretching vibrational states in HOCl[46] and H_2O[47] found the single (local) mode picture to hold up to $v = 4$. Furthermore, based on the comparison of the VCI calculations at different theory levels and the convergence tests with the size of the harmonic basis, it is also unlikely that the multidimensional VCI results are reliable for $v = 4$. Neither of the two approaches, therefore, appears well-suited for computing molecular properties for highly excited vibrational states ($v > 3$).

Interaction and mixing of the individual normal modes is only one of several potential problems associated with the single mode approximation. Even in the complete absence of intermode coupling, this treatment neglects the contribution of the zero point energy (ZPE) states of the remaining (anharmonic) modes to the property of interest, in this case the dipole moment. In order to estimate this effect, we keep the assumption of the noninteracting, separable normal modes, which allows writing the total wave function as a product of the one-dimensional functions

$$|\Psi(R_1, R_2, ..., R_N)\rangle = |\psi_{v_1}(R_1)\psi_{v_2}(R_2)...\psi_{v_N}(R_N)\rangle \tag{13}$$

where R_i is the normal coordinate (non-mass-weighted), as above. The applicability of the single mode approximation for calculation of the dipole moment also requires a separable dipole moment operator

$$\mu(R_1, R_2, ..., R_N) = \mu_1(R_1) + \mu_2(R_2) + ... + \mu_N(R_N) \tag{14}$$

i.e., all derivatives of the type $(\partial^2\mu/\partial R_i \partial R_j)$ in the Taylor expansion of the dipole moment are negligible for $i \neq j$. This, to some extent, resembles the "second harmonic approximation" used in calculation of absorption intensities in the (harmonic) vibrational spectra. The dipole moment expansion, however, is not truncated at the linear term; only the cross derivatives are neglected:

$$\mu_\alpha = \mu_{0,\alpha} + \sum_i \left(\frac{\partial \mu_{i,\alpha}}{\partial R_i} R_i + \frac{1}{2} \frac{\partial^2 \mu_{i,\alpha}}{\partial R_i^2} R_i^2 + \frac{1}{3!} \frac{\partial^3 \mu_{i,\alpha}}{\partial R_i^3} R_i^3 + ... \right)$$

$$= \mu_{0,\alpha} + \sum_i (\mu_{i,\alpha} - \mu_{0,\alpha}) \tag{15}$$

where $\alpha = \{x, y, z\}$ and μ_0 is the minimum energy structure (equilibrium) dipole moment. Then the ZPE-corrected dipole moment for the ν_3 (C═O stretch) vibrational state v is

$$\mu_\alpha(\nu_3) = \mu_{\alpha,0} + \langle \psi_v(R_{\nu_3}) | \mu_{\nu_3,\alpha} - \mu_{0,\alpha} | \psi_v(R_{\nu_3}) \rangle + \sum_{i \neq \nu_3} \langle \psi_0(R_i) | \mu_{i,\alpha} - \mu_{0,\alpha} | \psi_0(R_i) \rangle \tag{16}$$

Under these assumptions, the contribution of the additional normal modes can be evaluated using only 1D potential energy surfaces for each mode. It must be stressed, however, that assumptions analogous to those made for ν_3 are not automatically valid for all normal modes. Especially the highly anharmonic, low frequency vibrations may not be separable.

For this "ZPE-corrected" single mode approximation, we have calculated the potential energy and dipole moments as functions of normal coordinate displacements for all nine normal modes of FA at the BPW91/cc-pVDZ level (Supporting Information, Table S4). In analogy to the computations of the vibrationally averaged dipole moments for the ν_3 as described in Computational Methods, the vibrational

wave functions for each mode can be obtained and substituted into eq 16. In Table 4, we compare the vibrationally averaged BPW91/cc-pVDZ dipole moments computed for the local ν_3 mode only (eq 6, Table 1) with those including the ZPE correction for all vibrational modes (eq 16).

The ZPE contribution of the remaining modes causes a noticeable decrease in the ν_3 vibrationally averaged dipole moments by approximately 0.03 D for all states. The shift towards lower dipole moment values, although relatively small ($\gtrsim$2%), is qualitatively consistent with the BPW91/cc-pVDZ VCI anharmonic results, for convenience also shown in Table 4. The ground vibrational state dipole moment for the ZPE-corrected single mode is just ~0.015 D greater than the VCI anharmonic result, but also more than 0.03 D lower than the experimental dipole. The differences with respect to the VCI calculations increase for the excited states due to a steeper dependence of the single mode dipole moment on the vibrational state (Figures 2 and 4), which is obviously unaffected by the correction. The $v = 1$ dipole moment is also lower than the experiment, but the error is approximately equal to the noncorrected single mode calculation.

The ZPE contribution also accounts, at least in part, for one of the discrepancies between the single mode and VCI calculations as to whether the vibrationally averaged dipole moment for the ground ν_3 state is smaller (VCI calculations) or greater (single mode calculations) than the equilibrium (minimum energy) structure. The ZPE-corrected single mode dipole moment value becomes lower than that computed for the minimum energy structure (1.4026 D).

We note that the single mode treatment of all vibrational modes also gives, within this approximation, the vibrational frequencies and intensities of the fundamental and overtone transitions for all nine vibrational modes of FA. These results are presented in the Supporting Information (Tables S10 and S11) and compared with the experimental data (as well as with the VCI calculations) to provide additional tests for the performance of the single mode approximation at the BPW91/cc-pVDZ level. In addition, vibrationally averaged dipole moments for the excited vibrational states of all vibrational modes can be obtained. For reference, we present the computed dipole moments for the first excited vibrational states ($v = 1$) (Supporting Information, Table S4) along with the full anharmonic VCI results (Supporting Information, Table S5).

Our calculations have also provided tests for the performance of different levels of theory, including DFT, post-HF correlated wave function methods, and a number of basis sets. The errors in the computed molecular properties at different levels of theory as compared to experiment are reflected in both single mode and VCI anharmonic calculations, and some systematic trends can be inferred. For example, with the same basis sets, the post-HF dipole moments are uniformly smaller than those from the DFT. These trends are already apparent from the dipole moments within the harmonic approximation. As a consequence, with smaller basis sets the DFT yields dipole moments closer to experiment, while the post-HF ones are too low. With larger basis, however, in particular the correlation consistent basis sets augmented with diffuse functions, the calculated dipole moment values systematically increase and the post-HF predictions improve relatively to the DFT ones, which become much too large.

One of the interesting results was a surprisingly good overall performance of the BPW91/cc-pVDZ level. We have used this particular level of theory previously in simulations of the vibrational amide I (predominantly amide C═O stretch) spectra of model amides, since it appeared to give the best agreement of the harmonic vibrational frequency with the experimental gas-phase value for *N*-methylacetamide (NMA).[20−22] In this study, within the single mode approximation, we obtained the best agreement between BPW91/cc-pVDZ calculations with experiment for the dipole moments as well as the frequencies and IR intensities (Supporting Information). In the anharmonic VCI calculations, the agreement of both the predicted frequencies and dipole moments with experiment is worse, but still remarkably good in comparison with the other levels tested.

The unusually good performance of BPW91/cc-pVDZ must be due to a fortuitous cancellation of errors of the density functional and the basis set. The results obtained with the BPW91 systematically differ from those computed with the other methods: e.g., the harmonic vibrational frequencies are systematically lower (Supporting Information, Table S6). On the other hand, the cc-pVDZ basis set with all the computational methods produces higher vibrational frequencies than any other basis except 6-31G(d,p). These errors seem to compensate for each other very well in the combination BPW91/cc-pVDZ. With larger basis sets BPW91 tends to underestimate the vibrational frequencies, while for the other methods as the computed frequencies become lower with increasing basis set size, the agreement with experiment improves. For dipole moments, the BPW91 results are not as dramatically different from the other methods but fall in between the higher B3LYP values and the too low post-HF ones. The cc-pVDZ basis, however, again gives systematically the lowest dipole moments. While MP2 with this basis yields dipole moments that are much too small compared to the experiment and B3LYP/cc-pVDZ dipoles are systematically too high, the BPW91/cc-pVDZ-calculated dipole again falls very close to the experimental value. Thus, on an empirical basis, the BPW91/cc-pVDZ level of theory appears as a useful and computationally cheap model for the vibrational properties of the carbonyl group.

5. Conclusion

The explanation of multiple closely spaced peaks of the *trans*-FA ν_3 band in low-temperature, solid pH_2 required a reliable estimate of dipole moments in different ν_3 vibrational states. Calculations of vibrationally averaged molecular dipole moments represent a challenge, in particular for higher vibrationally excited vibrational states. We have tested two different methodologies: the single mode treatment and multidimensional limited VCI calculations. With both methods, very good results for the ground and first excited state dipole moments were obtained; however, the reliability of the predictions for higher excited states is difficult to verify. The linearity of the dipole moment as a function of

vibrational energy quanta, independently obtained from the VCI calculations at two DFT levels, suggests that the single mode picture is valid up to the third excited state ($v = 3$). Apart from the breakdown of the single mode picture due to combination with other normal modes and/or nonseparability of the dipole operator, the single mode approximation also neglects the zero point contribution of the remaining modes to the vibrationally averaged dipole moment. This contribution is non-negligible and the resulting dipole moments more closely correspond to those obtained from VCI. Unfortunately, even the VCI calculations are unlikely to be reliable for highly excited vibrational state properties, due in part to the anharmonic corrections based on the truncated Taylor series and in part to the truncation of the VCI harmonic basis, dictated by the computational cost of the VCI calculation. Of the different computational levels tested, we found the BPW91/cc-pVDZ to give remarkably good overall performance. In general, DFT results, while quite close with smaller basis sets, seem to depart increasingly from the experiment upon increasing the basis set size. In contrast, post-HF methods, while requiring large basis sets, appear to converge to better agreement with the available experimental data. A more detailed evaluation of the performance of different computational methods remains difficult due to the scarcity of the dipole moment experimental data. However, the results demonstrate that the computational methodology provides robust tools for studies of the molecular electrostatic properties in various vibrational states as well as interactions with the environment. In particular, the predicted vibrationally averaged dipole moments well-explained the quadrupole-dipole splitting of the IR lines of the FA in the solid pH_2 matrix.

Acknowledgment. This work was supported in part by the National Science Foundation CAREER: 0846140 grant to J.K., Grant Agency of the Czech Republic (202/07/0732, to P.B.), and the Grant Agency of the Academy of Sciences (A400550702, M200550902, to P.B. and J.K.).

Supporting Information Available: Anharmonic vibrationally averaged structural parameters, convergence of VCI dipole moments with the size of harmonic basis, potential energy surfaces and dipole moment functions for additional levels of theory, single mode and VCI vibrational frequencies and absorption intensities, dipole moments for all FA normal modes in single mode approximation, and VCI dipole moments for singly excited FA vibrational states. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Stone, A. J. *The Theory of Intermolecular Forces (International Series of Monographs on Chemistry)*; Oxford University Press: New York, 1997; pp 36−49.

(2) Craig, D. P.; Thirunamachandran, T. *Molecular quantum electrodynamics*; Dover Publications: New York, 1998; pp 142−182.

(3) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486.

(4) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 3rd ed.; Wiley-VCH: New York, 2003; pp 329−388.

(5) Nigam, S.; Rutan, S. *Appl. Spectrosc.* **2001**, *55*, 362A.

(6) Redington, R. L. *J. Mol. Spectrosc.* **1977**, *65*, 171.

(7) Marèchal, Y. *J. Chem. Phys.* **1987**, *87*, 6344.

(8) Weber, W. H.; Maker, P. D.; Johns, J. W. C.; Weinberger, E. *J. Mol. Spectrosc.* **1987**, *121*, 243.

(9) Hurtmans, D.; Herregodts, F.; Herman, M.; Lievin, J.; Campargue, A.; Garnache, A.; Kachanov, A. A. *J. Chem. Phys.* **2000**, *113*, 1535.

(10) Freytes, M.; Hurtmans, D.; Kassi, S.; Lievin, J.; Auwera, J. V.; Campargue, A.; Herman, M. *Chem. Phys.* **2002**, *283*, 47.

(11) Lundell, J.; Rasanen, M.; Latajka, Z. *Chem. Phys.* **1994**, *189*, 245.

(12) Pettersson, M.; Lundell, J.; Khriachtchev, L.; Rasanen, M. *J. Am. Chem. Soc.* **1997**, *119*, 11715.

(13) Madeja, F.; Markwick, P.; Havenith, M.; Nauta, K.; Miller, R. E. *J. Chem. Phys.* **2002**, *116*, 2870.

(14) Macoas, E. M. S.; Lundell, J.; Pettersson, M.; Khriachtchev, L.; Fausto, R.; Rasanen, M. *J. Mol. Spectrosc.* **2003**, *219*, 70.

(15) Marushkevich, K.; Khriachtchev, L.; Rasanen, M. *J. Chem. Phys.* **2007**, *126*, 241102.

(16) Marushkevich, K.; Khriachtchev, L.; Rasanen, M. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5748.

(17) Olbert-Majkut, A.; Ahokas, J.; Lundell, J.; Pettersson, M. *Chem. Phys. Lett.* **2009**, *468*, 176.

(18) Paulson, L. O.; Anderson, D. T. *J. Phys. Chem. A* **2009**, *113*, 1770.

(19) Nyquist, R. A. *Appl. Spectrosc.* **1986**, *40*, 336.

(20) Kubelka, J.; Keiderling, T. A. *J. Phys. Chem. A* **2001**, *105*, 10922.

(21) Amunson, K. E.; Kubelka, J. *J. Phys. Chem. B* **2007**, *111*, 9993.

(22) Ackels, L.; Stawski, P.; Amunson, K. E.; Kubelka, J. *Vib. Spectrosc.* **2008**, *50*, 2.

(23) Bouř, P. *J. Chem. Phys.* **2004**, *121*, 7545.

(24) Andrushchenko, V.; Matějka, P.; Anderson, D. T.; Kaminský, J.; Horníček, J.; Paulson, L. O.; Bouř, P. *J. Phys. Chem. A* **2009**, *113*, 9727.

(25) Daněček, P.; Bouř, P. *J. Comput. Chem.* **2007**, *28*, 1617.

(26) Takahashi, K.; Sugawara, M.; Yabushita, S. *J. Phys. Chem. A* **2002**, *106*, 2676.

(27) Wallace, R. *Chem. Phys.* **1975**, *11*, 189.

(28) Swofford, R. L.; Long, M. E.; Albrecht, A. C. *J. Chem. Phys.* **1976**, *65*, 179.

(29) Takahashi, K.; Sugawara, M.; Yabushita, S. *J. Phys. Chem. A* **2003**, *107*, 11092.

(30) Takahashi, K.; Sugawara, M.; Yabushita, S. *J. Phys. Chem. A* **2005**, *109*, 4242.

(31) Henry, B. R. *Acc. Chem. Res.* **1977**, *10*, 207.

(32) Lemus, R.; Frank, A. *Int. J. Quantum Chem.* **1999**, *75*, 465.

(33) Jensen, P. *Mol. Phys.* **2000**, *98*, 1253.

(34) Daněček, P.; Kapitán, J.; Baumruk, V.; Bednárová, L.; Kopecký, V.; Bouř, P. *J. Chem. Phys.* **2007**, *126*, 224513.

(35) Dračínský, M.; Kaminský, J.; Bouř, P. *J. Chem. Phys.* **2009**, *130*, 094106.

(36) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J., J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98; Revision A.6*; Gaussian, Inc.: Pittsburgh PA, 1998.

(37) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

(38) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669.

(39) Wallace, B. A.; Meyer, H. *J. Low Temp. Phys.* **1974**, *15*, 297.

(40) ACESII - Mainz-Austin-Budapest version: Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J., with contributions from Auer, A. A.; Bernholdt, D. E.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; Price, D. R.; Ruud, K.; Schiffmann, F. Tajti, A.; Varner, M. E.; Vázquez, J. and the integral packages: MOLECULE (J. Almlöf and P. R. Taylor), PROPS (Taylor, P. R.), and ABACUS (Helgaker, T.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J.).

(41) Cerjan, C.; Kulander, K. C. *Comput. Phys. Commun.* **1991**, *63*, 529.

(42) Sugawara, M.; Kato, M.; Fujimura, Y. *Chem. Phys. Lett.* **1991**, *184*, 203.

(43) Bouř, P. *J. Phys. Chem.* **1994**, *98*, 8862.

(44) Bouř, P.; Bednárová, L. *J. Phys. Chem.* **1995**, *99*, 5961.

(45) Kapitán, J.; Hecht, L.; Bouř, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1003.

(46) Callegari, A.; Theule, P.; Schmied, R.; Rizzo, T. R.; Muenter, J. S. *J. Mol. Spectrosc.* **2003**, *221*, 116.

(47) Shostak, S. L.; Ebenstein, W. L.; Muenter, J. S. *J. Chem. Phys.* **1991**, *94*, 5875.

(48) Shostak, S. L.; Muenter, J. S. *J. Chem. Phys.* **1991**, *94*, 5883.

CT900608T

JCTC Journal of Chemical Theory and Computation

Accurate Harmonic/Anharmonic Vibrational Frequencies for Open-Shell Systems: Performances of the B3LYP/N07D Model for Semirigid Free Radicals Benchmarked by CCSD(T) Computations

Cristina Puzzarini,*,† Malgorzata Biczysko,‡ and Vincenzo Barone§

Dipartimento di Chimica "G. Ciamician", Università di Bologna, Via F. Selmi 2, 40126 Bologna, Italy, Dipartimento di Chimica "Paolo Corradini" and CR-INSTM Village, Università di Napoli Federico II, Complesso Univ. Monte S. Angelo, via Cintia, 80126 Napoli, Italy, and Scuola Normale Superiore, piazza dei Cavalieri 7, 56126 Pisa, Italy

Received November 10, 2009

Abstract: Impressive growth of computer facilities and effective implementation of very accurate quantum mechanical methods allow, nowadays, the determination of structures and vibrational characteristics for small- to medium-sized molecules to a very high accuracy. Since the situation is much less clear for open-shell species, we decided to build a suitable database of harmonic and anharmonic frequencies for small-sized free radicals containing atoms of the first two rows of the periodic table. The level of theory employed is the CCSD(T) model in conjunction with triple- and quadruple-ζ basis sets, whose accuracy has been checked with respect to the available experimental data and/or converged quantum mechanical computations. Next, in view of studies of larger open-shell systems, we have validated the B3LYP/N07D model with reference to the above database: our results confirm previous suggestions about the remarkable reliability and reduced computational cost of this computational method. A number of test computations show that basis set extension has negligible effects and other density functionals (including last generation ones) deliver significantly worse results. Increased accuracy can be obtained, instead, by using CCSD(T) harmonic frequencies and B3LYP/N07D anharmonic corrections.

1. Introduction

Computational chemistry experiments have already been proven to provide highly accurate results for small molecules,[1−5] clearly demonstrating their potentiality as key tools for the prediction and understanding of spectroscopic properties of all kinds of molecular systems. At present, thanks to the progress in hardware and software, the a priori prediction of accurate low-lying vibrational levels of semirigid polyatomic molecules is becoming a viable task. It is now widely recognized that the computation of semidiagonal quartic force fields at the CCSD(T) (Coupled Clusters with Single, Double, and perturbative inclusion of Triple excitations)[6] level in conjunction with sufficiently large basis sets (at least of triple-ζ quality) followed by an effective second-order perturbative treatment usually provides results with an accuracy on the order of 10−15 cm^{-1} for fundamental transitions.[7−19] Although perturbative vibrational treatment remains highly cost-effective for quite large systems, the unfavorable scaling of the CCSD(T) model with the number of active electrons limits the determination of quartic force fields to molecules containing at most five to six atoms. Additionally, a simple reduction of computational cost by combining correlated quantum mechanical (QM) methods with a small basis set should not be recommended, due to the quite unpredictable accuracy of the results. Thus,

* To whom correspondence should be addressed. E-mail: cristina.puzzarini@unibo.it.

† Università di Bologna.

‡ Università di Napoli Federico II.

§ Scuola Normale Superiore.

10.1021/ct900594h
Published on Web 02/04/2010

extension of computational studies to larger systems requires cheaper yet reliable electronic structure approaches.

Recently, several authors have reported anharmonic force fields for small- and medium-sized semirigid molecules computed by methods rooted in density functional theory (DFT).[20−25] Among the functionals tested, the hybrid ones provide satisfactory results when coupled to basis sets of at least double-ζ plus polarization quality supplemented by diffuse sp functions. An even more effective approach in terms of good accuracy, obtained at a computationally reduced cost, is based on the additivity of DFT anharmonic corrections to CCSD(T) harmonic force fields. This is well-known to further improve the agreement with experimental data.[26−30] The situation is much more involved for open-shell systems since experimental data are scarce and often of questionable reliability. Recently, we decided to start a comprehensive research project to extend the spectroscopic accuracy for studies of open-shell systems. To this purpose, some of the results available in the literature in this framework (HCO, HOC, HSiO, HOSi, HCS, HSC, HCCO, H_2CN, F_2CN, and F_2BO radicals) will be used here to build a representative set of reference force fields. Furthermore, a number of additional radicals have been purposely investigated for the present work (FCO, FSiO, FCS, FSC, NH_2, PH_2, $HOCl^+$, H_2BO, Cl_2CN, NH_3^+, PH_3^+, H_2CO^+, HCNN, and HNCN) in order to enlarge the benchmark set dimensions.

The present work has a 3-fold aim: (*i*) harmonic frequency results obtained by means of the B3LYP/N07D computational model, recently proposed for spectroscopic studies of free radicals, are validated by the comparison with corresponding computations at the CCSD(T) level; (*ii*) anharmonic corrections to vibrational frequencies computed from CCSD(T) quartic force fields are employed to confirm good performance of the B3LYP/N07D model in evaluating anharmonic contributions; (*iii*) the performances of last generation density functionals and basis set convergence are tested for both harmonic and anharmonic frequencies on a reduced set of radicals. Although, in order to facilitate comparison with CCSD(T) results, we have limited our benchmark set to the tri- and tetratomic radicals, we consider it significant, as molecules containing first- as well as second-row elements are included.

The manuscript is organized as follows. In the next section, all the computational details concerning the CCSD(T) and DFT calculations as well as the methodology employed are described. Then, the results are presented and discussed in the frame of the aims presented above.

2. Methodology and Computational Details

2.1. Coupled-Cluster Computations. The coupled-cluster (CC) computations have been performed at the CC level of theory with single and double excitations augmented by a quasi-perturbative account for triple substitutions [CCSD(T)].[6] The adequacy of the coupled-cluster model to treat the systems under consideration has been checked by means of the T_1[31,32] and D_1 (ROCCSD)[33] diagnostics, which provide values lower than 0.025 and 0.045, respectively, for all radicals considered. All CCSD(T) computations have been carried out in conjunction with the correlation-consistent polarized (aug)-cc-pV*n*Z, with n = T and Q, basis sets,[34,35] in the frozen core (fc) approximation. The CFOUR program package[36] has been employed, and the unrestricted Hartree−Fock (UHF) wave function has been used as a reference in the CCSD(T) computations.

Once the required geometry optimizations have been performed at the level of theory considered; anharmonic force field calculations have been carried out only for the main isotopic species of each radical under study. The harmonic part of the force field has been obtained using analytic second derivatives of the energy,[37] and the corresponding anharmonic force field has been determined in a normal coordinate representation via numerical differentiation of the analytically evaluated force constants as described in refs 11 and 38. Subsequently, the force field has been used to compute spectroscopic parameters, such as anharmonic frequencies, by means of the vibrational second-order vibrational perturbation theory (VPT2),[39] as implemented in the CFOUR package.[36]

2.2. Density Functional Theory Computations. Density functional theory computations have been performed with the B3LYP/N07D model, defined by a combination of the well-known B3LYP[40] density functional with the recently developed polarized double-ζ N07D basis set,[41−43] properly tailored for studying free radicals. This basis set has been constructed by adding a reduced number of polarization and diffuse functions to the 6-31G set (see refs 41 and 42 for details), leading to an optimum compromise between reliability and computer time.

All structures have been optimized using tight convergence criteria followed by computations of anharmonic frequencies by means of the VPT2 approach,[39,44] as implemented in the Gaussian package.[45] Semidiagonal quartic force fields have been evaluated by the numerical differentiation (with the standard 0.025 Å step) of analytical second derivatives. As VPT2 computations are sensitive to the proper treatment of Fermi resonances, it is crucial to automatically neglect nearly singular contributions (deperturbed computations). This is performed by effectively removing interactions in the second-order treatment, which are more properly treated in the first order. For this purpose, the VPT2 implementation[44] makes use of the criteria proposed by Martin and Boese,[23] implemented in an automated scheme that has already been shown to provide accurate results for at least fundamental bands.[46]

In addition to computations with the B3LYP functional, known to provide reliable predictions of vibrational properties,[20,21] it seemed important to check also the performance of other density functionals, in view of the recent developments aiming at improvement of long-range effects, i.e., the proper description of dispersion interaction or of the properties of excited electronic states. Concerning spectroscopic studies, an unsatisfactory description of vibrational frequencies computed by the M05(6)-2X[47,48] or LC-ωPBE[49] functionals has been found in the study of adenine,[50] in contrast to the accurate results obtained for functionals originating from the B3LYP one, such as B3LYP-D(M)[51,52] or CAM-B3LYP.[53] In this context, we decided to extend our benchmark to a broader range of recently introduced density

functionals, namely, M06,[48] the wB97 family,[54,55] and HSE06,[56] to check their performance in evaluating vibrational properties of open-shell systems. For the sake of completeness, the parameter-free PBE0 functional[57] has also been taken into consideration, and the basis set effect has been accounted for by comparison with the aug-cc-pVTZ[35] results.

All DFT computations have been performed employing the Gaussian suite of programs for quantum chemistry.[45]

2.3. Computations with Hybrid CCSD(T)/DFT Approach. As already mentioned in the Introduction, the hybrid CCSD(T)/DFT approach has also been applied to evaluate anharmonic frequencies. Such a methodology takes into account that CCSD(T) computations can be prohibitively expensive, even for medium-sized systems, whenever extended samplings of potential energy surfaces (PES) are required, whereas the main discrepancies between anharmonic frequencies computed at the CCSD(T) and B3LYP levels are related to the inaccuracies within the harmonic part.[26−28] In this respect, the hybrid CCSD(T)/DFT scheme[26−30] stands as a viable route to extend predictions of accurate anharmonic frequencies to relatively large systems.

In the present case, two possible approaches have been implemented. In the most simple one (DPT2), the harmonic frequencies computed at the CCSD(T) level are *a posteriori* corrected by anharmonic contributions ($\Delta\nu$) derived from VPT2 computations performed at the DFT level: $\nu_{\mathrm{CCSD(T)/DFT}} = \omega_{\mathrm{CCSD(T)}} + \Delta\nu_{\mathrm{DFT}}$. Such an approximation has already been validated for several closed and open shell systems (see for instance refs 26−29). The second option is based on the introduction of the harmonic frequencies evaluated at the CCSD(T) level into the VPT2 computations along with the 3rd and 4th force constants obtained at the DFT level. Such an approach is available in the Gaussian package through the InDerAU and InFreq options and might significantly improve the quality of the results in difficult cases, i.e., when large discrepancies between harmonic frequencies computed at the DFT and CCSD(T) levels or Fermi resonances take place.

3. Results and Discussion

3.1. Equilibrium Structure. Evaluation of harmonic as well as anharmonic frequencies implies accurate prediction of molecular structures, which need to be computed at the corresponding level of theory. Therefore, in view of the connection between molecular structure and vibrational frequencies, we start our discussion by assessing the accuracy of the geometry parameters obtained at the computational levels subsequently applied to frequency evaluation. The results are organized as follows. In Table 1, optimized geometries for those radicals for which a systematic basis-set investigation at the CCSD(T) level is available are reported, while in Table 2 the structures computed at the CCSD(T) and B3LYP levels are compared to the available experimental data. For all the other radicals, molecular structures obtained at the CCSD(T) and B3LYP levels are collected in the Supporting Information.

Before proceeding in this discussion, general notes on the accuracy of CCSD(T) and B3LYP optimized geometries are deserved. Concerning coupled-cluster determinations, according to ref 58, the accuracy of molecular structures of closed-shell systems optimized at the CCSD(T) level in conjunction with the (aug)-cc-pVQZ basis sets is expected to be on the order of 0.002−0.004 Å for distances and 0.1−0.3° for angles. This accuracy range is not straightforwardly applicable to open-shell molecules as for them the situation is more involved from a theoretical as well as experimental point of view. The accuracy of CCSD(T)/(aug)-cc-pVQZ optimized geometries can be derived from the comparison to computed geometries for which extrapolation to the complete basis set limit (CBS) and core correlation (CV) effects are accounted for. This comparison is summarized for selected radicals in Table 1. On the basis of the results available in the literature[5,15,26,27,59,60] and those of Table 1, we can conclude that the overall accuracy is analogous to that of closed-shell systems. Going a little more into detail, we note that quadruple-ζ bases tend to provide geometrical parameters quite close to the CBS limit, in particular when only first-row elements are involved. The discrepancies are in general in the ranges 0.001−0.01 Å for distances and 0.01−0.1° for angles, where the larger values refer to the parameters involving second-row atoms. CV corrections range from 0.001 to 0.005 Å for bonds and from 0.01 to 0.1° for angles. Once again, larger values are observed when second-row elements are involved.

Concerning the accuracy of DFT structures, in benchmark studies performed to validate and develop the B3LYP/N07D computational model, it has been pointed out that the overall performance of the N07D basis set is comparable to that of the aug-cc-pVDZ one,[41] but with increased computational efficiency. Additionally, recent B3LYP/N07D studies performed on vinyl,[61] propyl, and phenyl radicals[62] showed agreements within 0.01 Å for C−C and C−H bond lengths, and 2° for angles with respect to highly accurate computational studies at the CCSD(T)[63,64] and multireference[64] levels in conjunction with extended basis sets. Further hints on the accuracy of geometric structures computed at the B3LYP/N07D level of theory can be obtained from the analysis of the data reported in Tables 1 and 2. Comparison with CCSD(T)/VQZ results shows that in most cases DFT slightly overestimates bond lengths by about 0.005−0.02 Å, while larger discrepancies, 0.026 and 0.034 Å, have been found for two difficult cases: the C−S and Si−O bonds, respectively. In contrast, C−N and B−O bond lengths are slightly (less than 0.01 Å) underestimated by DFT, whereas all angles agree within 1.5°. In view of the good accuracy of the CCSD(T)/VQZ structures, similar conclusions can be drawn from the comparison of B3LYP/N07D structures with most elaborated computational methodologies. In summary, B3LYP/N07D optimized geometries differ by about 0.01 Å and 1° from their CCSD(T) counterparts. Thus, B3LYP/N07D structures can be considered sufficiently accurate for studies not requiring extreme accuracy, and this is particularly encouraging for large systems for which expensive coupled cluster calculations are still unfeasible.

Table 1. Basis-Set Effects on Equilibrium Geometrical Parameters Computed at the CCSD(T) Level[a]

molec./param.	B3LYP/N07D	CCSD(T)/VQZ	CCSD(T)/V5Z	CCSD(T)/V6Z	CCSD(T)/CBS	CCSD(T)/CBS+CV
HCS[b]						
R(CH)	1.0942	1.0877	1.0872	1.0872	1.0872	1.0854
R(CS)	1.5783	1.5660	1.5622	1.5610	1.5594	1.5554
∠HCS	131.61	131.99	132.21	132.24	132.28	132.49
HSC[b]						
R(SH)	1.3850	1.3671	1.3662	1.3662	1.3662	1.3639
R(CS)	1.6773	1.6517	1.6457	1.6442	1.6421	1.6370
∠HSC	103.72	102.89	103.01	103.04	103.08	103.06
NH_2[c]						
R(NH)	1.0312	1.0250	1.0248	1.0247	1.0246	1.0233
∠HNH	103.55	102.72	102.95	103.01	103.08	103.21
PH_2[c]						
R(PH)	1.4358	1.4186	1.4183	1.4183	1.4183	1.4155
∠HPH	91.7	91.87	91.88	91.88	91.88	91.81
H_2CN[d]						
R(CH)	1.1016	1.0947	1.0946	1.0946	1.0946	1.0933
R(CN)	1.2479	1.2493	1.2486	1.2482	1.2479	1.2451
∠HCN	121.78	121.16	121.11	121.10	121.09	121.10
H_2BO						
R(BH)	1.2132	1.2046	1.2046	1.2046	1.2046	1.2025
R(BO)	1.2861	1.2927	1.2915	1.2911	1.2909	1.2868
∠HBO	120.33	118.99	119.00	119.00	119.00	119.08
F_2CN[e]		(aug)				+(diff)
R(CF)	1.3177	1.3079 (1.3087)	1.3076	1.3074	1.3074	1.3056
R(CN)	1.2572	1.2565 (1.2570)	1.2560	1.2558	1.2557	1.2532
∠FCN	124.55	124.51 (124.52)	124.52	124.52	124.52	124.5
F_2BO[e]		(aug)				+(diff)
R(BF)	1.3311	1.3133 (1.3142)	1.3129	1.3128	1.3125	1.3102
R(BO)	1.3602	1.3640 (1.3648)	1.3636	1.3634	1.3632	1.3602
∠FBO	119.4	119.30 (119.29)	119.29	119.29	119.29	119.28
Cl_2CN						
R(CCl)	1.7508	1.7354	1.7242	1.7236	1.7232	1.7195
R(CN)	1.2458	1.2493	1.2491	1.2489	1.2488	1.2474
∠ClCN	122.04	122.00	121.88	121.88	121.88	121.88
NH_3^+[c]						
R(NH)	1.0281	1.0212	1.0210	1.0210	1.0210	1.0201
∠HNH	120.0	120.0	120.0	120.0	120.0	120.0
PH_3^+[c]						
R(PH)	1.4124	1.3974	1.3972	1.3972	1.3972	1.3946
∠HPH	113.15	113.21	113.21	113.21	113.21	113.24

[a] DFT results are given for comparison. [b] Ref 15. [c] Ref 70. [d] Ref 5. [e] Ref 27.

As far as the comparison with experiment is concerned, the results of Table 2 allow us to point out that equilibrium structures obtained at the CCSD(T) level usually well agree with the experimental ones when the latter are accurately determined. On the other hand, in most cases, the accuracy and reliability of experimental determinations for open-shell species are quite limited; for such cases, best estimates (CBS+CV) such as those reported in Table 1 represent better reference geometries. We furthermore note that the B3LYP/N07D level of theory is able to provide a good semiquantitative description of equilibrium structures, with geometrical parameters usually overestimated by about 0.01−0.02 Å for bond lengths. In view of this, we do not expect any particular effect of molecular structures on vibrational frequencies. We only note that larger discrepancies have been observed for FSC, HSiO, and FSiO, for which at the DFT level the F−S and F−Si distances are overestimated by about 0.1 Å, the Si−O bond length is longer by more than 0.03 Å, and the ∠HSiO and ∠FSiO angles are underestimated by ∼4 and ∼2°, respectively. For these molecules, we actually noticed for some vibrational modes discrepancies in harmonic frequencies larger than the average (i.e., on the order of 100 cm^{-1}), but a clear correspondence between anomalies in the molecular structure and large deviations for vibrational frequencies cannot be drawn.

3.2. Vibrational Frequencies. We start the analysis of vibrational frequencies by discussing the accuracy of harmonic frequencies computed by different methods as well as that of the corresponding anharmonic contributions. This is then followed by the validation of the computational approaches considered in this work against available experimental data.

Concerning harmonic frequencies computed at the CCSD(T) level, it is well-known from the literature that for closed-shell systems they have an overall accuracy of 15−20 cm^{-1} when basis sets of at least triple-ζ quality are used (see for example refs 19 and 65). For radicals, the situation is not so well assessed, but some recent investigations seem to confirm that an analogous accuracy can be reached (see for instance refs 5, 26, and 27). Hints on the quality of CCSD(T) anharmonic frequencies can be derived from Table 3, where they are compared with the available experimental data. From this comparison, it is apparent that, on average, anharmonic frequencies computed at the CCSD(T)/cc-pVQZ level for the harmonic part and employing either a triple-ζ or a quadruple-ζ basis

Table 2. Equilibrium Geometric Parameters Computed at the CCSD(T) and DFT Levels of Theory for Selected Tri- and Tetra-Atomic Free Radicals

	B3LYP/N07D	CCSD(T)/VnZ	exptl
HCO		$n = Q$	r_e[a]
R(CH)	1.1274	1.1162	1.1191(50)
R(CO)	1.1804	1.1759	1.1754(15)
∠HCO	124.36	124.57	124.43(25)
HSiO		$n = Q$	r_e[b]
R(SiH)	1.5393	1.5209	1.4971(fix)
R(SiO)	1.5672	1.5331	1.5286(2)
∠HSiO	115.84	119.63	116.8(1)
HCS		$n = Q$	r_0[c]
R(CH)	1.0942	1.0877	1.079(3)
R(CS)	1.5783	1.5660	1.56228(3)
∠HCS	131.61	131.99	132.8(3)
HSC		$n = Q$	r_e[d]
R(SH)	1.3850	1.3671	1.379(3)
R(CS)	1.6773	1.6517	1.6343(5)
∠HSC	103.72	102.89	104.2(2)
NH_2		$n = Q$	r_e[e]
R(NH)	1.0312	1.0250	1.0254 (12)
∠HNH	103.55	102.72	102.85 (14)
H_2CN		$n = Q$	r_0[f]
R(CH)	1.1016	1.0947	1.11(postulated)
R(CN)	1.2479	1.2493	1.247
∠HCN	121.78	121.16	121.65
F_2CN		$n = Q$ (aug)	r_0[g]
R(CF)	1.3177	1.3087	1.31(postulated)
R(CN)	1.2572	1.2570	1.265
∠FCN	124.55	124.52	123.25
F_2BO		$n = Q$ (aug)	r_0[i]
R(BF)	1.3311	1.3142	1.30(5)
R(BO)	1.3602	1.3648	1.40(5)
∠FBO	119.4	119.29	126(5)
NH_3^+		$n = Q$	r_e[k]
R(NH)	1.0281	1.0212	1.014
∠HNH	120.0	120.0	120.0

[a] Ref 71. [b] Ref 72. [c] Ref 73. [d] Ref 74. [e] Ref 75. [f] Ref 76. [g] Ref 77. [i] Ref 78. [k] Ref 79.

for the anharmonic contributions are able to reproduce experimental data within 10−20 cm^{-1}. Some larger deviations, i.e., more than 40 cm^{-1}, are observed only for C−H stretching modes: ν_1 of HCO and H_2CO^+ and ν_5 of H_2CN. Interestingly, in the two former cases the B3LYP/N07D computations led to significantly better agreement with respect to the experiment. For the ν_5 vibration of H_2CN, the discrepancy should be attributed to difficulties in assignment of the experimental band, as already discussed in refs 5 and 66, rather than to possible anharmonic resonances, as supported by the good agreement between variational and perturbative computations.[66−68] However, it should be noted that, for molecular systems which are significantly plagued by resonances, improvements with respect to perturbative results are expected by performing fully variational frequency computations.

The good performance of the B3LYP model for the computation of harmonic frequencies of open-shell species has already been pointed out,[69] and data collected in Table 3 are in line with such findings. Indeed, the comparison between harmonic frequencies computed at the CCSD(T) and B3LYP levels shows that the latter are on average off by only 1.5% from the CCSD(T) reference, with a maximum error of about 4.5%. Nevertheless, such an overall good agreement leads in absolute terms to discrepancies in the range of −30 to +60 cm^{-1}, which implies that the B3LYP/N70D level of theory might not be adequate for accurate spectroscopic studies. More interestingly, both methods predict very similar anharmonic corrections which agree within about 10 cm^{-1}. The only exceptions are FCO and H_2CO^+, but in both cases the largest discrepancies are related to the incorrect prediction of Fermi resonances at the DFT level. It should be stressed that, in the VPT2 implementation employed, vibrational modes which might be involved in Fermi or Darling-Denison resonances are excluded from the perturbative treatment[44] and are thus treated separately at the variational level. However, as such mode selection is performed on the basis of harmonic frequencies, any inaccuracy in the latter strongly influences the final results. In fact, as soon as accurate harmonic frequencies are included in the VPT2 treatment (and also used to predetermine resonances), significant improvements are achieved.

The first conclusion that can be drawn is that the good accuracy of anharmonic corrections computed by DFT validates the use of a hybrid scheme in which the harmonic part is computed with high accuracy by means of coupled cluster theory while anharmonic contributions are obtained by relatively inexpensive DFT computations. As mentioned before, the standard approach is to simply add DFT anharmonic contributions, $\Delta\nu$, to harmonic frequencies computed at the CCSD(T) level; however, sometimes the second approach, where harmonic CCSD(T) frequencies are directly used in perturbative vibrational analysis along with 3rd and 4th force constants computed at the DFT level, might be more suitable.[28] This is indeed the case for FCO and H_2CO^+, where application of corrected harmonic values leads to much better agreement (within 3 cm^{-1} for FCO and 15 cm^{-1} for H_2CO^+) between the hybrid scheme and full CCSD(T) results. It seems, therefore, that hybrid computations using CCSD(T) quadratic force fields and B3LYP/N07D cubic and semidiagonal quartic force constants closely approach the accuracy of complete CCSD(T) anharmonic computations at a much reduced computational cost. Anharmonic frequencies derived from full B3LYP/N07D force fields also agree fairly well with the experiment, but in most cases the improvements achieved by the hybrid CCSD(T)/cc-pVQZ//B3LYP/N07D scheme are significant: for example, for F_2CN, the mean absolute error (MAE) reduces from 18 cm^{-1} to 6 cm^{-1}. The hybrid scheme is thus advisable whenever possible.

In the following analysis, anharmonic DFT and hybrid frequencies will be compared with benchmark CCSD(T) values which have been shown to yield results in good agreement with the experiment. Tables 4 and 5 list such results for the triatomic and tetratomic radicals not considered in Table 3, respectively. From these tables, the conclusions drawn from Table 3 are confirmed: with only a few exceptions (some of them already discussed in the previous section), the anharmonic B3LYP/N07D frequencies agree with their CCSD(T) counterparts within 5−60 cm^{-1}. All the discrepancies are then removed once the hybrid approach is considered; in fact, in almost all cases, the agreement is within 1−30 cm^{-1}. For the

Table 3. Harmonic (ω) and Anharmonic (ν) Vibrational Frequencies (in cm^{-1}) Computed at Various Levels of Theory for Selected Tri- and Tetra-Atomic Free Radicals

	B3LYP/N07D		CCSD(T)[a]		CCSD(T)+DFT[b]	
	ω	ν	ω	ν	ν	exptl[c]
HCO			CBS/aCV	CBS+QZ		
ν_1	2677	2441	2717	2460	2481	2435
ν_2	1930	1906	1905	1878	1882	1868
ν_3	1099	1066	1120	1093	1087	1081
FCS			augVQZ	aVTZ		
ν_1	1306	1275	1338	1308	1306	1297
ν_2	934	919	933	919	918	918
ν_3	456	451	461	456	456	457
FCO			augVQZ	augVTZ		
ν_1	1915	1847	1900	1864	1862[d]	1862
ν_2	1022	977	1054	1025	1028[d]	1026
ν_3	621	610	634	624	624[d]	628[e]
NH_2			VQZ	VQZ		
ν_1	3346	3193	3377	3219	3223	3219
ν_2	1518	1475	1549	1504	1505	1497[f]
ν_3	3450	3281	3471	3297	3301	3301
PH_2			VQZ	VQZ		
ν_1	2353	2254	2401	2305	2302	2298
ν_2	1120	1096	1127	1101	1103	1102
ν_3	2364	2262	2409	2310	2306	
$HOCl^+$			VQZ	VQZ+augVTZ		
ν_1	3463	3276	3524	3338	3337	
ν_2	1257	1221	1288	1248	1251	
ν_3	888	875	915	901	902	830 ± 50
HCCO			VQZ	VQZ+VTZ		
ν_1	3348	3217	3352	3214	3221	3232[g]
ν_2	2085	2047	2097	2056	2058	2022[h]
ν_3	1258	1250	1246	1245	1239	
ν_4	546	553	567	576	574	
ν_5	497	447	505	467	455	494[i]
ν_6	488	517	500	524	529	
H_2CN			VQZ	VQZ		
ν_1	2985	2822	3000	2836	2837	2820[j]
ν_2	1720	1692	1706	1672	1677	1725[k]
ν_3	1384	1346	1383	1343	1346	1330
ν_4	988	968	989	967	969	950
ν_5	3044	2867	3069	2893	2893	3103[k]
ν_6	931	912	939	913	919	909
F_2CN			augVQZ	aVQZ+augVTZ		
ν_1	1790	1759	1811	1781	1781	1771
ν_2	957	941	974	957	958	955
ν_3	540	535	552	546	546	
ν_4	666	658	679	673	671	660
ν_5	1239	1206	1295	1262	1263	1257
ν_6	492	487	501	496	497	497
NH_3^+			VQZ	VQZ		
ν_1	3351	3206	3375	3231	3230	3232[l]
ν_2	873	914	865	910	906	917[m]
ν_3	3538	3365	3559	3388	3386	
ν_4	3538	3365	3559	3388	3386	3389[n]
ν_5	1532	1492	1551	1507	1510	1507[o]
ν_6	1532	1491	1551	1507	1510	
PH_3^+			VQZ	VQZ		
ν_1	2448	2361	2497	2400	2410	
ν_2	726	652	751	670	678	695[p]
ν_3	2525	2435	2568	2469	2477	
ν_4	2525	2436	2568	2469	2479	2462
ν_5	1040	1018	1054	1029	1032	1044
ν_6	1040	1017	1054	1029	1031	
H_2CO^+			VQZ	VQZ		
ν_1	2783	2603	2807	2625	2635[d]	2580
ν_2	1711	1742	1676	1636	1652[d]	1675
ν_3	1251	1207	1263	1210	1220[d]	1210[r]
ν_4	1062	1037	1068	1039	1044[d]	1036[s]
ν_5	2873	2675	2915	2711	2724[d]	2718[r]
ν_6	848	858	848	823	810[d]	823[s]

[a] CCSD(T) data from HCO/HOC, ref 80; HCCO, ref 81; H_2CN/F_2CN, ref 5; NH_3^+/PH_3^+, ref 70; others, this work. [b] Anahrmonic corrections at the B3LYP/N07D level. [c] Experimental results from references: HCO,[82] FCS,[83] FCO,[84] NH_2,[85] PH_2,[86] $HOCl^+$, [87] H_2CN,[88] F_2CN,[77] PH_3^+,[89] H_2CO^+.[90,91] [d] Anahrmonic corrections computed with CCSD(T) harmonic frequency and 3rd and 4th force constants obtained at the B3LYP/N07D level. [e] Ref 92. [f] Ref 93. [g] Ref 94. [h] Ref 95. [i] Ref 96. [j] Ref 97. [k] Ref 98. [l] Refs 99, 100 and 101. [m] Ref 102. [n] Ref 79. [o] Refs 103, 104, 100, and 105. [p] Also, ref 106. [r] Also, refs 107 and 108.

molecules gathered in Tables 4 and 5, the simple DPT2 hybrid scheme usually provided satisfactory results, whereas, for more demanding cases like HOC, FSiO, and Cl_2CN radicals, good accuracy has only been achieved with the InFreq approach. For the latter radical, in analogy to what was discussed above for FCO and H_2CO^+, the

Table 4. Anharmonic Frequencies (in cm^{-1}) for Triatomic Radicals Obtained with B3LYP/N07D, CCSD(T), and Hybrid Models

	B3LYP/N07D	CCSD(T)	CCSD(T)+DFT[a]
HOC		VQZ	
ν_1	2887	3144	3117[b]
ν_2	1337	1375	1374[b]
ν_3	1075	1108	1109[b]
HSiO		V5Z(PES)//VQZ	
ν_1	1774	1828//1829	1847//1842
ν_2	1068	1166// 1168	1162//1161
ν_3	558	622// 622	624//627
HOSi		V5Z(PES)//VQZ	
ν_1	3634	3667//3667	3671//3671
ν_2	797	869//869	860//859
ν_3	749	743//749	755//760
HCS		CBS(PES)//VQZ	
ν_1	2962	2993//2992	2979//2977
ν_2	1175	1188//1181	1187//1180
ν_3	807	794//802	791//798
HSC		CBS(PES)//VQZ	
ν_1	2196	2252//2256	2264//2266
ν_2	881	930//920	931//921
ν_3	745	773//768	785//779
FSC		aVTZ	
ν_1	1030	1046	1046
ν_2	518	626	622
ν_3	214	241	242
FSiO		aVTZ	
ν_1	1064	1206	1181[b]
ν_2	758	845	844[b]
ν_3	265	308	301[b]

[a] Anharmonic corrections at the B3LYP/N07D level. [b] Anharmonic corrections computed with the CCSD(T) harmonic frequency and 3rd and 4th force constants obtained at the B3LYP/N07D level.

difficulties are related to Fermi resonances. On the other hand, for HOC and FSiO, abnormally large discrepancies (over 100 cm^{-1}) between harmonic frequencies computed at the CCSD(T) and DFT levels are an issue.

The last part of our work is devoted to the analysis of the performance of several recently developed density functionals for computation of vibrational frequencies within both the DFT/N07D and hybrid models. For this purpose, Table 6 compares mean absolute errors (MAEs), with respect to CCSD(T) computations, over all normal modes of H_2CN, N_2H^+, and FCS radicals, for which a good agreement with both the experiment and the benchmark CCSD(T) studies has been obtained at the B3LYP/N07D level. First of all, we discuss the overall accuracy of harmonic frequencies obtained by means of the functionals under study. It has to be noted that an accuracy only slightly lower than that of B3LYP, with discrepancies on the order of 15−25 cm^{-1} with respect to CCSD(T) results, has been achieved by most functionals, except LC-ωPBE and wB97(X), which show MAEs in range 40−50 cm^{-1}. Such findings are also valid for anharmonic frequencies, for which most functionals show MAEs lower than 30 cm^{-1}. However, larger discrepancies are observed for higher frequencies, as depicted in panel a of Figure 1. Additionally, in two cases, namely, for LC-ωPBE and M06-2X, the MAE exceeded 60 cm^{-1}: in the former case, the error originates from the inaccurate harmonic frequency values, but in the latter case the problems are directly related to anharmonic corrections. All functionals have been also tested for their performance within hybrid models, and the results obtained applying both hybrid approaches (the simple DPT2 correction and the InFreq) are listed in Table 6. Additional insights can also be drawn from panel b of Figure 1, which shows differences with respect to CCSD(T) results as a function of the frequency for all normal modes of the selected radicals. It is immediately apparent that among the tested functionals those originated

Table 5. Anharmonic Frequencies (in cm^{-1}) for Tetratomic Radicals Obtained with B3LYP/N07D, CCSD(T), and Hybrid Models

	B3LYP/N07D	CCSD(T)	CCSD(T)+DFT[a]
HCNN		VQZ+VTZ	
ν_1	3013	3020	3015
ν_2	1828	1746	1742
ν_3	1233	1176	1176
ν_4	830	861	871
ν_5	533	530	534
ν_6	485	519	518
HNCN		VQZ+VTZ	
ν_1	3304	3313	3308
ν_2	1881	1781	1773
ν_3	1188	1131	1137
ν_4	1019	1037	1030
ν_5	491	472	473
ν_6	470	461	465
H_2BO		VQZ+VTZ	
ν_1	2312	2327	2331
ν_2	1384	1355	1354
ν_3	949	975	988
ν_4	881	912	900
ν_5	2323	2364	2368
ν_6	545	547	564
F_2BO		aVQZ+VTZ	
ν_1	1372	1427	1425
ν_2	842	873	873
ν_3	442	462	460
ν_4	648	662	659
ν_5	1330	1427	1423
ν_6	362	389	387
Cl_2CN		VQZ+VTZ	
ν_1	1710	1577	1575[b]
ν_2	564	589	583[b]
ν_3	363	364	364[b]
ν_4	495	492	490[b]
ν_5	776	828	837[b]
ν_6	281	283	283[b]

[a] Anharmonic corrections at the B3LYP/N07D level. [b] Anharmonic corrections computed with CCSD(T) harmonic frequency and 3rd and 4th force constants obtained at the B3LYP/N07D level.

Table 6. Performances of Modern Density Functionals in the DFT/N07D Models[a]

	DFT/N07D		$CCSD(T)_{harm} + DFT_{anh}$	
MAE [cm^{-1}]	Harm	Anh	DPT2	InFreq
B3LYP	15.0	14.5	2.4	4.1
B3LYP-D	19.9	18.7	4.0	4.3
CAM-B3LYP	20.0	24.3	4.4	3.0
PBE0	19.6	22.3	4.9	3.6
LC-ωPBE	50.9	61.1	11.5	6.5
M06	25.2	23.2	13.9	15.3
M06-2X	25.5	74.9	54.0	52.4
wB97	39.0	31.4	18.6	23.5
wB97X	33.7	27.1	12.4	15.7
wB97XD	21.9	26.5	17.3	20.0
HSE06	17.6	20.9	6.0	4.7

[a] Mean absolute error (MAE) with respect to CCSD(T) computations over all normal modes of H_2CN, NH_2^+, and FCS.

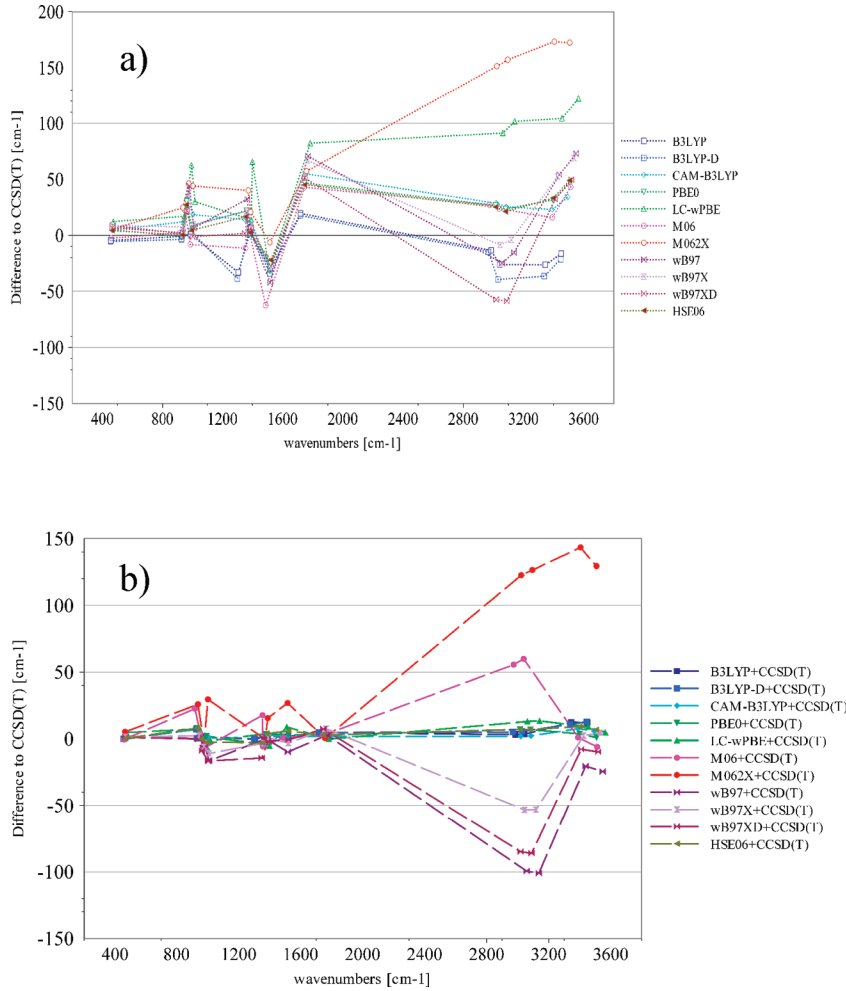


Figure 1. Performance of different density functionals in prediction of vibrational frequencies beyond harmonic approximation. (a) Perturbative computations performed fully at the DFT level. (b) Hybrid CCSD(T)+DFT scheme. Relative discrepancies from the values computed at the CCSD(T) level are shown for each normal mode of H_2CN, NH_2^+, and FCS, which are listed according to their wavenumbers.

from B3LYP and PBE0, together with the recently introduced HSE06, yield accurate anharmonic frequencies, with MAEs as low as 3−7 cm^{-1}. It should be noted that, in the case of the LC-ωPBE model, such good agreement can be achieved only by the InFreq approach, due to large errors in harmonic frequencies. In contrast, recently developed functionals belonging to the M06 and wB97 families yield significantly less accurate results, with MAEs in the range 12−55 cm^{-1}, and with discrepancies up to 150 cm^{-1}, thus far off the accuracy required for spectroscopic studies. It is also noteworthy that the performance of M06 functionals is significantly worsened by the addition of Hartree−Fock exchange (2X). A further check of the above conclusions has been performed by comparing frequencies obtained using DFT/N07D models with much more expensive DFT/aug-cc-pVTZ computations. The results are listed in Table 7 for the H_2CN radical showing that indeed the large errors in anharmonic corrections obtained at the M06/N07D and wB97/N07D levels are not removed by the larger basis set.

Table 7. Harmonic Frequencies and Anharmonic Contributions Computed by Several Modern Density Functionals with N07D and aug-cc-pVTZ Basis Sets[a]

	harmonic frequencies		anharmonic contribution	
MAE [cm^{-1}]	N07D	aVTZ	N07D	aVTZ
B3LYP	10.4	20.4	2.9	3.5
B3LYP-D	16.5	26.6	6.7	7.4
CAM-B3LYP	19.7	15.8	5.6	6.3
PBE0	14.7	16.5	6.6	7.2
LC-ωPBE	43.8	29.9	12.3	12.7
M06	24.0	36.4	26.0	20.6
M06-2X	20.7	12.7	55.5	66.2
wB97	35.9	28.9	31.4	23.0
wB97X	31.1	22.7	21.4	16.5
wB97XD	17.8	11.0	30.5	24.7
HSE06	13.5	16.0	6.7	6.2

[a] Mean absolute errors (MAE) with respect to benchmark results from CCSD(T) computations for the H_2CN radical.

4. Conclusions

The present paper has been devoted to the validation of the DFT/N07D and hybrid CCSD(T)/DFT models for studying vibrational properties of free radicals. In this respect, we have chosen several small radicals containing first- and second-row elements for which it was possible to compare methodologies rooted in the density functional theory with computations at the CCSD(T) level. At the same time, comparison with several experimental results allowed ex-

ploration of the ability of the DFT/N07D and CCSD(T)/DFT models to critically analyze results from spectroscopic studies. It has been shown that the accuracy of the CCSD(T) model is nearly the same for closed- and open-shell systems: this finding is by itself significant since the reliability of experimental data is usually strongly reduced when going from closed-shell to open-shell systems, essentially because of the short lifetime of the latter ones. On the other hand, the B3LYP/N07D model is fairly robust for geometries and harmonic and anharmonic frequencies and becomes nearly quantitative when used only for anharmonic contributions to be added to harmonic force fields obtained at more sophisticated levels. Furthermore, extension of the N07D basis set has only marginal effects, and the use of other functionals (including the most recent ones) does not improve the results. As a matter of fact, some of the most successful last generation functionals (M06-2X and wB97X) provide quite disappointing results: this suggests that vibrational frequencies should be added to the databases used for the optimization of parameters in this kind of functionals. In conclusion, we think that the B3LYP/N07D model can represent a very effective tool coupling a remarkable reliability in the computation of geometric and vibrational properties of organic and inorganic free radicals with a very favorable scaling with the number of electrons. Furthermore, empirical dispersion terms (leading to B3LYP-D) can be added without worsening the performances of the model whenever dispersion interactions come into play. The relatively low computational cost of the B3LYP/N07D computational model allows taking into proper account the vibrational effects beyond the harmonic approximation even for quite large systems of biological and/or technological interest.

Acknowledgment. This work was supported by Italian MIUR (PRIN 2006), CNR (PROMO 2006), and by the University of Bologna (RFO funds). The large scale computer facilities of the VILLAGE network (http://village.unina.it) and the Wroclaw Centre for Networking and Supercomputing are acknowledged for providing computer resources. The authors also thank the CINECA supercomputing center for a grant of computer time on the IBM SP5 machine.

Supporting Information Available: Geometry parameters computed at the B3LYP/N07D and CCSD(T)/VnZ levels for HOC, HOSi, FCS, FSC, FCO, FSiO, $HOCl^+$, HCCO and H_2CO^+ radicals. Harmonic and anharmonic frequencies computed by DFT/N07D, DFT/aug-cc-pVTZ, and hybrid models for the H_2CN radical, as well as harmonic and anharmonic frequencies computed by DFT/N07D and hybrid models for NH_2 and FCS radicals. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Jensen, P.; Bunker, P. R. *Computational Molecular Spectroscopy*; John Wiley & Sons: United Kingdom, 2000.

(2) Carter, S.; Handy, N. C.; Puzzarini, C.; Tarroni, R.; Palmieri, P. *Mol. Phys.* **2000**, *98*, 1697–1712.

(3) Biczysko, M.; Tarroni, R.; Carter, S. *J. Chem. Phys.* **2003**, *119*, 4197–4203.

(4) Puzzarini, C.; Barone, V. *Chem. Phys. Lett.* **2008**, *462*, 49–52.

(5) Puzzarini, C.; Barone, V. *Chem. Phys. Lett.* **2009**, *467*, 276–280.

(6) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(7) East, A. L. L.; Allen, W. D.; Klippenstein, S. J. *J. Chem. Phys.* **1995**, *102*, 8506–8532.

(8) Martin, J. M. L.; Lee, T. J.; Taylor, P. R.; Francois, J.-P. *J. Chem. Phys.* **1995**, *103*, 2589–2602.

(9) Dateo, C. E.; Lee, T. J. *Spectrochim. Acta A* **1997**, *53*, 1065–1077.

(10) Breidung, J.; Thiel, W. *Theor. Chem. Acc.* **1998**, *100*, 183–190.

(11) Stanton, J. F.; Lopreore, C. L.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 7190–7196.

(12) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 9218–9220.

(13) Breidung, J.; Thiel, W.; Gaus, J.; Stanton, J. F. *J. Chem. Phys.* **1999**, *110*, 3687–3696.

(14) Ruden, T.; Taylor, P. R.; Helgaker, T. *J. Chem. Phys.* **2003**, *119*, 1951–1960.

(15) Puzzarini, C. *J. Chem. Phys.* **2005**, *123*, 024313/1–14.

(16) Bizzocchi, L.; Degli Esposti, C.; Puzzarini, C. *Mol. Phys.* **2006**, *104*, 2627–2640.

(17) Puzzarini, C. *J. Mol. Spectrosc.* **2007**, *242*, 70–75.

(18) Baldacci, A.; Stoppa, P.; Pietropolli Charmet, A.; Giorgianni, S.; Cazzoli, G.; Puzzarini, C.; Larsen, R. W. *J. Phys. Chem. A* **2007**, *111*, 7090–7097.

(19) Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. *J. Phys. Chem. A* **2007**, *111*, 11242–11248.

(20) Barone, V. *J. Phys. Chem. A* **2004**, *108*, 4146–4150.

(21) Barone, V. *Chem. Phys. Lett.* **2004**, *383*, 528–532.

(22) Burcl, R.; Handy, N. C.; Carter, S. *Spectrochim. Acta A* **2003**, *59*, 1881–1893.

(23) Boese, A. D.; Martin, J. *J. Phys. Chem. A* **2004**, *108*, 3085–3096.

(24) Cané, E.; Miani, A.; Trombetti, A. *J. Phys. Chem. A* **2007**, *111*, 8218–8222.

(25) Cané, E.; Trombetti, A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2428–2432.

(26) Puzzarini, C.; Barone, V. *J. Chem. Phys.* **2008**, *129*, 084306/1–7.

(27) Puzzarini, C.; Barone, V. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6991–6997.

(28) Carbonniere, P.; Lucca, T.; Pouchan, C.; Rega, N.; Barone, V. *J. Comput. Chem.* **2005**, *26*, 384–388.

(29) Begue, D.; Carbonniere, P.; Pouchan, C. *J. Phys. Chem. A* **2005**, *109*, 4611–4616.

(30) Begue, D.; Benidar, A.; Pouchan, C. *Chem. Phys. Lett.* **2006**, *430*, 215–220.

(31) Lee, T. J.; Taylor, P. R. *Int. J. Quant. Chem. Symp.* **1989**, *23*, 199–207.

(32) Lee, T. J.; Scuseria, G. E. *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Kluwer: Dordrecht, The Netherlands, 1995; p 47.

(33) Leininger, M. L.; Nielsen, I. M. B.; Crawford, T. D.; Janssen, C. L. *Chem. Phys. Lett.* **2000**, *328*, 431–436.

(34) Dunning, T. H., J. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(35) Kendall, A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(36) Stanton, J. F.; Gauss, J.; Harding, M. E.; Szalay, P. G. *CFOUR*, a quantum chemical program package, with contributions from Auer, A. A.; Bartlett, R. J.; Benedikt, U.; Berger, C.; Bernholdt, D. E.; Bomble, Y. J.; Christiansen, O.; Heckert, M.; Heun, O.; Huber, C.; Jagau, T.-C.; Jonsson, D.; Jusélius, J.; Klein, K.; Lauderdale, W. J.; Matthews, D.; Metzroth, T.; O'Neill, D. P.; Price, D. R.; Prochnow, E.; Ruud, K.; Schiffmann, F.; Stopkowicz, S.; Varner, M. E.; Vázquez, J.; Watts, J. D.; and Wang, F. and the integral packages MOLECULE (J. Almlöf and P. R. Taylor), PROPS (P. R. Taylor), ABACUS (T. Helgaker, H. J. Aa. Jensen, P. Jørgensen, and J. Olsen), and ECP routines by A. V. Mitin and C. van Wüllen, development version. For the current public version, see http://www.cfour.de (accesed Jan 13, 2010).

(37) Gauss, J.; Stanton, J. F. *Chem. Phys. Lett.* **1997**, *276*, 70–77.

(38) Stanton, J. F.; Gauss, J. *Int. Rev. Phys. Chem.* **2000**, *19*, 61–95.

(39) Mills, I. M. *Molecular Spectroscopy: Modern Research*; Academic: New York, 1972.

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

(41) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751–764.

(42) Barone, V.; Cimino, P. *Chem. Phys. Lett.* **2008**, *454*, 139–143.

(43) Barone, V.; Cimino, P. *J. Chem. Theory Comput.* **2009**, *5*, 192–199.

(44) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108/1–10.

(45) 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.; 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.; 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.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fo, D. J. *Gaussian Development Version*, revision H.01; Gaussian, Inc.: Wallingford, CT, 2006.

(46) Barone, V.; Festa, G.; Grandi, A.; Rega, N.; Sanna, N. *Chem. Phys. Lett.* **2004**, *388*, 279–283.

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

(48) Zhao, Y.; Truhlar, D. G. *Theor. Chim. Acta* **2008**, *120*, 215–241.

(49) Jacquemin, D.; Perpète, E.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105/1–12.

(50) Biczysko, M.; Panek, P.; Barone, V. *Chem. Phys. Lett.* **2009**, *475*, 105–110.

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

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

(53) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(54) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106/1–15.

(55) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(56) Henderson, T.; Izmaylov, A. F.; Scalmani, G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 044108/1–9.

(57) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(58) Helgaker, T.; Jørgensen, P.; Olsen, J. *Electronic-Structure Theory*; Wiley: Chichester, U. K., 2000.

(59) Puzzarini, C. *Chem. Phys.* **2008**, *346*, 45–52.

(60) Puzzarini, C.; Barone, V. *J. Chem. Theory Comput.* **2009**, *5*, 2378–2387.

(61) Barone, V.; Bloino, J.; Biczysko, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1092−1101.

(62) Barone, V.; Biczysko, M.; Cimino, P. In *Carbon-Centered Free Radicals and Radical Cations*; Forbes, M. D. E., Ed.; John Willey & Sons, Inc.: New York, 2010; Chapter: Interplay of stereo electronic vibrational and environmental effects in tuning physico-chemical properties of carbon centered radicals, pp 105−139.

(63) Koziol, L.; Levchenko, S. V.; Krylov, A. I. *J. Phys. Chem. A* **2006**, *110*, 2746–2758.

(64) Mebel, A. M.; Chen, Y.-T.; Lin, S.-H. *Chem. Phys. Lett.* **1997**, *275*, 19–27.

(65) Begue, D.; Carbonniere, P.; Barone, V.; Pouchan, C. *Chem. Phys. Lett.* **2005**, *416*, 206–211.

(66) Barone, V.; Carbonniere, P.; Pouchan, C. *J. Chem. Phys.* **2005**, *122*, 224308/1–8.

(67) Pouchan, C.; Carbonniere, P. *Chin. J. Chem. Phys.* **2009**, *22*, 123–128.

(68) Carbonniere, P.; Pouchan, C. *Int. J. Quantum Chem.* **2010**, *110*, 578–585.

(69) Byrd, E. F. C.; Sherrill, C. D.; Head-Gordon, M. *J. Phys. Chem. A* **2001**, *105*, 9736–9747.

(70) Puzzarini, C. *Theor. Chem. Acc.* **2008**, *120*, 325–336.

(71) Hirota, E. *J. Mol. Struct.* **1986**, *146*, 237–252.

(72) Izuha, M.; Yamamoyo, S.; Saito, S. *J. Mol. Struct.* **1997**, *413−414*, 527–535.

(73) Habara, H.; Yamamoyo, S.; Amano, T. *J. Chem. Phys.* **2002**, *116*, 9232–9238.

(74) Habara, H.; Yamamoyo, S. *J. Chem. Phys.* **2000**, *112*, 10905–10911.

(75) Kobayashi, K.; Ozeki, H.; Saito, S.; Tonooka, M.; Yamamoyo, S. *J. Chem. Phys.* **1997**, *107*, 9289–9296.

(76) Yamamoyo, S.; Saito, S. *J. Chem. Phys.* **1992**, *96*, 4157–4162.

(77) Jacox, M.; Milligan, D. *J. Chem. Phys.* **1968**, *48*, 4040–4046.

(78) Mathews, C. W. *J. Mol. Spectrosc.* **1966**, *19*, 203–223.

(79) Bawendi, M.; Rehfuss, B.; Dinelli, B.; Okumura, M.; Oka, T. *J. Chem. Phys.* **1989**, *90*, 5910–5917.

(80) Marenich, A. V.; Boggs, J. E. *J. Phys. Chem. A* **2003**, *107*, 2343–2350.

(81) Szalay, P. G.; Tajti, A.; Stanton, J. F. *Mol. Phys.* **2005**, *103*, 2159–2168.

(82) Sappey, A.; Crosley, D. *J. Chem. Phys.* **1990**, *93*, 7601–7608.

(83) Caspary, N.; Wurfel, B.; Thoma, A.; Schallmoser, G. B. V. *Chem. Phys. Lett.* **1993**, *212*, 329–339.

(84) Nagai, K.; Yamada, C.; Endo, Y.; Hirota, E. *J. Mol. Spectrosc.* **1981**, *90*, 249–272.

(85) McKellar, A.; Vervloet, M.; Burkholder, J.; Howard, C. *J. Mol. Spectrosc.* **1990**, *142*, 319–335.

(86) Jakubek, Z.; Bunker, P.; Zachwieja, M.; Nakhate, S.; Simard, B.; Yurchenko, S.; Thiel, W.; Jensen, P. *J. Chem. Phys.* **2006**, *124*, 094306/1–5.

(87) Colbourne, D.; Frost, D.; McDowell, C.; Westwood, N. *J. Chem. Phys.* **1978**, *68*, 3574–3580.

(88) Pettersson, M.; Lundell, J.; Khriachtchev, L.; Rasanen, M. *J. Chem. Phys.* **1998**, *109*, 618–625.

(89) Yang, J.; Li, J.; Hao, Y.; Zhou, C.; Mo, Y. *J. Chem. Phys.* **2006**, *125*, 054311/1–12.

(90) Niu, B.; Shirley, D.; Bai, Y.; Daymo, E. *Chem. Phys. Lett.* **1993**, *201*, 212–216.

(91) Niu, B.; Shirley, D.; Bai, Y. *J. Chem. Phys.* **1993**, *98*, 4377–4390.

(92) Jacox, M. *J. Mol. Spectrosc.* **1980**, *80*, 257–271.

(93) Burkholder, J.; Howard, C.; McKellar, A. *J. Mol. Spectrosc.* **1988**, *127*, 415–424.

(94) Wilhelm, M.; McNavage, W.; Groller, R.; Dai, H.-L. *J. Chem. Phys.* **2008**, *128*, 064313/1–8.

(95) Unfried, K.; Curl, R. *J. Mol. Spectrosc.* **1991**, *150*, 86–98.

(96) Brock, L.; Mischler, B.; Rohlfing, E.; Bise, R.; Neumark, D. *J. Chem. Phys.* **1997**, *107*, 665–668.

(97) Cowles, D.; Travers, M.; Frueh, J.; Ellison, G. *J. Chem. Phys.* **1991**, *94*, 3517–3528.

(98) Jacox, M. *J. Phys. Chem.* **1987**, *91*, 6595–3660.

(99) Miller, P.; Colson, S.; Chupka, W. *Chem. Phys. Lett.* **1988**, *145*, 183–187.

(100) Dobber, M.; Buma, W.; de Lange, C. *J. Phys. Chem.* **1995**, *99*, 1671–1685.

(101) Bahng, M.-K.; Xing, X.; Baek, S.; Ng, C. *J. Chem. Phys.* **2005**, *123*, 084311/1–8.

(102) Thompson, W.; Jacox, M. *J. Chem. Phys.* **2001**, *114*, 4846–4854.

(103) Habenicht, W.; Reiser, G.; Muller-Dethlefs, K. *J. Chem. Phys.* **1991**, *95*, 4809–4820.

(104) Reiser, G.; Habenicht, W.; Muller-Dethlefs, K. *J. Chem. Phys.* **1993**, *98*, 8462–8468.

(105) Locht, R.; Leyh, B.; Hottmann, K.; Baumgartel, H. *Chem. Phys.* **1998**, *233*, 145–158.

(106) Maripuu, R.; Reineck, I.; Agren, H.; Nian-Zu, W.; Rong, J.; Veenhuizen, H.; Al-Shamma, S.; Karlsson, L.; Siegbahn, K. *Mol. Phys.* **1983**, *48*, 1255–1267.

(107) Liu, J.; Kim, H.-T.; Anderson, S. *J. Chem. Phys.* **2001**, *114*, 9797–9806.

(108) Schulenburg, A.; Meisinger, M.; Radi, P.; Merkt, F. *J. Mol. Spectrosc.* **2008**, *250*, 44–50.

CT900594H

JCTC Journal of Chemical Theory and Computation

Approximate Inclusion of Triple Excitations in Combined Coupled Cluster/Molecular Mechanics: Calculations of Electronic Excitation Energies in Solution for Acrolein, Water, Formamide, and *N*-Methylacetamide

Kristian Sneskov,*,† Eduard Matito,‡,† Jacob Kongsted,§ and Ove Christiansen†

The Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Aarhus, Langelandsgade 140, DK-8000 Aarhus C, Denmark, and Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Received December 1, 2009

Abstract: Electronic excitation energies are often significantly affected by perturbing surroundings such as, for example, solvent molecules. Correspondingly, for an accurate comparison between theory and experiment, the inclusion of solvent effects in high-level theoretical predictions is important. Here, we introduce the CCSDR(3)/MM model designed for an effective, flexible, and accurate prediction of electronic excitation energies in solution. The method is based on a hybrid coupled cluster/molecular mechanics (CC/MM) strategy including interactions between a solute described by CC methods and a solvent described by polarizable MM methods. The CCSDR(3)/MM includes triples effects in a computational tractable noniterative fashion. The resulting approach allows for both high-accuracy inclusion of triples effects and inclusion of solute−solvent interactions with polarization effects, as well as being applicable for averaging over many solvent configurations derived from, for example, molecular simulations. We test the proposed model using as a benchmark the two lowest-lying valence singlet excitations ($n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$) of acrolein, formamide, and *N*-methylacetamide in aqueous solution as well as liquid water, demonstrating how a systematic inclusion of many different effects leads to good agreement with experimental values. In doing so we also illustrate the theoretical challenges involved when investigating UV properties of solvated molecules.

1. Introduction

In traditional quantum chemistry the major focus is on obtaining accurate energies and properties of isolated molecules. However, facing the experimental demand of *solvated* molecules while simultaneously requiring a flexible description of the one- and N-electron space poses a formidable challenge to modern quantum chemistry. Brute-force large-scale macromolecular calculations using high-accuracy ab initio theory are currently unfeasible for the most popular methods. Therefore, many schemes have been suggested in order to incorporate the perturbing effects of the surroundings in an approximate fashion.[1] One successful approach is based on the combined quantum mechanics/molecular mechanics (QM/MM) scheme,[2,3] which is also the approach we follow in this study. The QM part is defined as the chemically most important region. Therefore, invoking the well-established CC approximation is very satisfactory from a theoretical point of view. Combined with molecular mechanics, CC/MM allows for an accurate calculation of molecular properties. The development of CC/MM and similar methods is an active research area,[4–9] and in this work we consider further development of a strategy that includes polarization interac-

* To whom correspondence should be addressed. E-mail: sneskov@chem.au.dk. Fax: +4586196199.

† University of Aarhus.

‡ Present address: Institute of Physics, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland.

§ University of Southern Denmark.

10.1021/ct900641w
Published on Web 02/01/2010

tion between the QM and MM systems in the context of electronic excitation energy calculations.

The development of systematic CC models is an essential tool for the continuing investigation of molecular properties. Especially, these models serve as a reference for the computationally less expensive density functional theory (DFT) methods. In a series of papers,[10,11] the CC hierarchy of models—CCS, CC2, CCSD, and CC3—was established and tested in benchmark calculations.[12] It was also shown in past publications[13] how a noniterative analogue to CC3, CCSDR(3), could be designed providing excitation energies of seemingly similar quality to the CC3 model. The CCSDR(3) model is more appealing from a computational perspective as it incorporates effects due to triply excited configurations (henceforward denoted triples effects) in a *noniterative* fashion. The usefulness of CCSDR(3) and its ability to give very similar excitation energies to CC3 has recently been confirmed by Sauer et al. in an extensive study of medium-sized organic molecules.[14] The CC3 and CCSDR(3) methods were derived in a response theory framework. Related frameworks include equation-of-motion CC (EOM-CC)[15] (providing CCSD excitation energies identical to CC response theory) and symmetry adapted cluster/configuration interaction (SAC-CI).[16] Over the years a number of other approximate methods for inclusion of higher excitations in coupled cluster calculations of electronic excitations have been suggested also within these frameworks; see the lists in refs 8, 15, 17–19. A few of these have been successfully combined with a point-charge description of the surroundings and applied for instance in a study of the excited states of uracil[8] as well as in a retinal proteins investigation.[20] Here, we maintain the focus on CCSDR(3) and introduce the CCSDR(3)/MM model designed especially for the calculation of excitation energies in solution, and we stress the inclusion of polarization effects in addition to point charges.

In order to rigorously compare with experiment, it is mandatory to account for the effect of dynamics. This is routinely done by molecular dynamics (MD) or Monte Carlo sampling techniques, where the effect of the surroundings (here a solvent) is accounted for by a force field. Extracting conformations allows for an approximate sampling of the actual molecular environment by ultimately averaging the properties at hand over the conformations in a suitable fashion. This averaging requires many single-point calculations, emphasizing further the need for efficiency and thus the attractiveness of a noniterative triples approach like CCSDR(3) relative to an iterative approach.

Recently, some of us published[21] an extensive study of *s-trans*-acrolein, where the two lowest-lying singlet excitations were categorized in great detail. In the present study we use the same molecular geometries as in the former investigation, allowing us to directly ascertain the effects of triples in the aforementioned excitations energies. As a further test of the model, we also calculate electronic excitation energies and oscillator strengths of two amides, formamide and *N*-methylacetamide (NMA), demonstrating in the process the challenges involved in both the theoretical assignment of valence excitation energies as well as the direct comparison with experiment. Finally, we also investigate liquid water formerly studied by some of us without inclusion of triples.[22,23]

This paper is organized in the following way: In section 2.1 we outline the CC approximation and the use of CC response theory to calculate vertical excitation energies, including a definition of an excitation energy in CCSDR(3). In section 2.2, we briefly review the inclusion of solvent effects in the CC methodology, ending the section by defining the CCSDR(3)/MM model. In section 3, we outline the computational details, while we present the calculated values for solvated acrolein, formamide, NMA, and liquid water in section 4. Finally, section 5 contains the concluding remarks.

2. Theory

2.1. Excitation Energies in Coupled Cluster Theory. In conventional CC theory for isolated molecules the energy and amplitude equations take the form

$$E_{\mathrm{CC}}(t) = \langle \mathrm{HF}|\hat{H}\exp(\hat{T})|\mathrm{HF}\rangle \tag{1}$$

$$e_{\mu_i}(t) = \langle \mu_i|\exp(-\hat{T})\hat{H}\exp(\hat{T})|\mathrm{HF}\rangle = 0 \tag{2}$$

where $|\mathrm{HF}\rangle$ is the Hartree−Fock reference state, $\hat{T} = \sum_{\mu_i} t_{\mu_i}\hat{\tau}_{\mu_i}$ is the cluster operator, and $\langle\mu_i|$ is the excitation projection manifold. In the CCSD approximation $i = 1, 2$, such that the cluster operator is approximated to include only singles and doubles excitations $\hat{T} = \hat{T}_1 + \hat{T}_2$. The basic CC approaches are now textbook material, and we refer to ref 24 for further detail.

In response theory, the excitation energies for the exact wave function are found as poles of the linear response function. Similarly for approximate CC wave functions, the excitation energies are found by solving the CC response eigenvalue equations either for the right eigenvector

$$\mathbf{A}\mathbf{R}_k = \omega_k\mathbf{R}_k \tag{3}$$

or for the left eigenvector

$$\mathbf{L}_k\mathbf{A} = \mathbf{L}_k\omega_k \tag{4}$$

Here $\mathbf{A}$ is the CC asymmetric Jacobian

$$A_{\mu_i\nu_j} = \langle \mu_i|\exp(-\hat{T})[\hat{H}, \hat{\tau}_{\nu_j}]|\mathrm{CC}\rangle \tag{5}$$

Using this definition of the Jacobian, we note that the identification of the eigenvalues of the CC Jacobian as the excitation energies holds also for intermediate models such as CC2 and CC3.[25] Choosing the excitation vectors to be biorthonormal, $\mathbf{L}_i\mathbf{R}_j = \delta_{ij}$, the excitation energy can be written as

$$\omega_k = \mathbf{L}_k\mathbf{A}\mathbf{R}_k \tag{6}$$

A careful analysis of the order of the excitation energies in the CCSD model (see, for example, ref 11) reveals that the singles- and doubles-dominated excitations are correct through second and first order in the ground state fluctuation potential, respectively. In the CC3 and CCSDR(3) models,[11,13] on the other hand, the most important (in a perturbational sense) triples contributions are included such that both

models yield excitation energies correct to third and second order for the single- and double-dominated excitations, respectively.

In short, the iterative CC3 model is defined by invoking two approximations compared to the full CCSDT model: (1) the form of the singles and doubles amplitude equations is retained while the triples amplitude equation is restricted to the terms entering in lowest nonvanishing order in the fluctuation potential (second order), and (2) the singles are treated as zeroth-order parameters, thus implicitly accounting for orbital relaxation due to an external potential giving good response functions.

The main drawback of CC3 is the iterative N^7-scaling step motivating the development of the noniterative analogue: CCSDR(3).[13] The CCSDR(3) excitation energy is defined as

$$\omega = \mathbf{L}^{\mathrm{SD}}\mathbf{A}^{\mathrm{SD}}(t_1^*, t_2^*)\mathbf{R}^{\mathrm{SD}} + \sum_{\mu_i, i=1,2} L^{\mathrm{SD}}_{\mu_i} \sum_{\nu_3} \times \frac{\langle \mu_i|[\tilde{U}^*, \hat{\tau}_{\nu_3}]|\mathrm{HF}\rangle\langle \nu_3|[\tilde{U}^*, \hat{R}_2^{\mathrm{SD}}] + [\tilde{U}^*, \hat{R}_1^{\mathrm{SD}}], \hat{T}_2^*|\mathrm{HF}\rangle}{\omega_{\nu_3} - \omega^{\mathrm{SD}}} + \sum_{\mu_2} L^{\mathrm{SD}}_{\mu_2}\langle \mu_2|[[\hat{U}, \hat{R}_1^{\mathrm{SD}}], \hat{T}_3^*]|\mathrm{HF}\rangle \quad (7)$$

Here ω^{SD} is the CCSD excitation energy while $\mathbf{L}^{\mathrm{SD}}$ and $\mathbf{R}^{\mathrm{SD}}$ are CCSD response eigenvectors and

$$\hat{R}_i^{\mathrm{SD}} = \sum_{\nu_i} R^{\mathrm{SD}}_{\nu_i}\hat{\tau}_{\nu_i} \quad \text{for } i = 1, 2 \quad (8)$$

In eq 7 we have implied a partitioning of the Hamiltonian into the Fock operator and the fluctuation potential $\hat{H} = \hat{F} + \hat{U}$.

$\tilde{U}$ is in turn a T_1 similarity transformed operator according to

$$\tilde{U} = \exp(-\hat{T}_1)\hat{U}\exp(\hat{T}_1) \quad (9)$$

$\mathbf{A}^{\mathrm{SD}}(t_1^*, t_2^*)$ is the CCSD Jacobian constructed with the triples-corrected amplitudes defined as

$$t^*_{\mu_i} = t^{\mathrm{SD}}_{\mu_i} + \frac{\langle \mu_i|[\tilde{H}, \hat{T}_3]|\mathrm{HF}\rangle}{\omega_{\mu_i}} \quad (10)$$

where $\hat{T}_3$ is constructed on-the-fly from the CCSD amplitudes according to

$$t_{\mu_3} = -\frac{\langle \mu_3|[\tilde{U}, \hat{T}_2]|\mathrm{HF}\rangle}{\omega_{\mu_3}} \quad (11)$$

where a canonical representation has been implied. ω_{μ_i} contains orbital energy differences between occupied and virtual orbitals. In the CCSDR(3) approximation, a one-step perturbative correction is applied while this same correction is performed until convergence in the CC3 model.

In passing, we note that the excitation energies are no longer found as poles of any response functions due to the perturbational nature of CCSDR(3).

2.2. Environmental Effects. In order to include the effects of a surrounding environment, in the present case a solvent, a set of interaction terms are augmented to the vacuum energy expression. This implies a partitioning of the terms into a vacuum and a solvent part. This partitioning carries over to the Jacobian such that the total Jacobian may be written as

$$\mathbf{A}^{\mathrm{tot}} = \mathbf{A}^{\mathrm{vac}} + \mathbf{A}^{\mathrm{solv}} \quad (12)$$

where the form of $\mathbf{A}^{\mathrm{vac}}$ is still given by eq 5, but using the in-solution amplitudes, while the solvent Jacobian is given by

$$A^{\mathrm{solv}}_{\mu_i\nu_j} = \langle \mu_i|\exp(-\hat{T})[\hat{T}^g, \hat{\tau}_{\nu_j}]|\mathrm{CC}\rangle + \langle \mu_i|\exp(-\hat{T})\hat{T}^{g\nu_j}|\mathrm{CC}\rangle \quad (13)$$

The effective operators $\hat{T}^g$ and $\hat{T}^{g\nu_j}$ are introduced as

$$\hat{T}^g = \sum_p \lambda_p\hat{X}_p + \sum_q \gamma_q(\langle \Lambda|\hat{Z}_q|\mathrm{CC}\rangle\hat{Y}_q + \langle \Lambda|\hat{Y}_q|\mathrm{CC}\rangle\hat{Z}_q) \quad (14)$$

$$\hat{T}^{g\nu_j} = \sum_q \gamma_q(\langle \Lambda|[\hat{Z}_q, \hat{\tau}_{\nu_j}]|\mathrm{CC}\rangle\hat{Y}_q + \langle \Lambda|[\hat{Y}_q, \hat{\tau}_{\nu_j}]|\mathrm{CC}\rangle\hat{Z}_q) \quad (15)$$

where we have applied the auxiliary state $\langle\Lambda| = (\langle \mathrm{HF}| + \sum_{i\mu_i}\bar{t}_{\mu_i}\langle \mu_i|)\exp(-\hat{T})$. The notation used here is very general and is used due to the flexibility, allowing for a simultaneous description of the CC/MM model as well as the more simplified dielectric continuum description of the surroundings, denoted the CC/DC model.[26] In the latter implicit description of the solvent, the solute is placed in a (spherical) cavity surrounded by a continuum described by a dielectric constant. This may in eqs 14 and 15 be represented by dropping the term with the summation over p and letting the sum over q represent a multipole expansion for the charge distribution of the solute, thus implying that in the case of CC/DC $\hat{Y}_q = \hat{Z}_q$ are related to multipole operators. The explicit CC/MM model describes the surroundings using a molecular mechanics force field such that the effective operators in eqs 14 and 15 describe the electrostatic interaction between solute and solvent. Thus, the sum over p is related to the partial charges distributed in the MM region while the effects of polarization are incorporated in the sum over q. In both cases concrete expressions of these operators can be found in ref 5. We emphasize that the appearance of the solvent Jacobian in eq 12 is a direct consequence of the explicit inclusion of polarization effects thorugh the polarizabilities. If no polarization is included (e.g., a simple point-charge model), no extra term in the Jacobian appears provided that the partial point-charges have been absorbed into the Hamiltonian, and it would be sufficient to use the in-solution amplitudes in the expression for the vacuum Jacobian. Clearly, this more widespread approximation leads to much simpler equations, but here it is not taken as default.

In a straightforward CC3 approach, one would iteratively introduce the effects of triples in both the description of the solute and the solute−solvent interaction terms. However, it is currently intractable to perform such high-accuracy

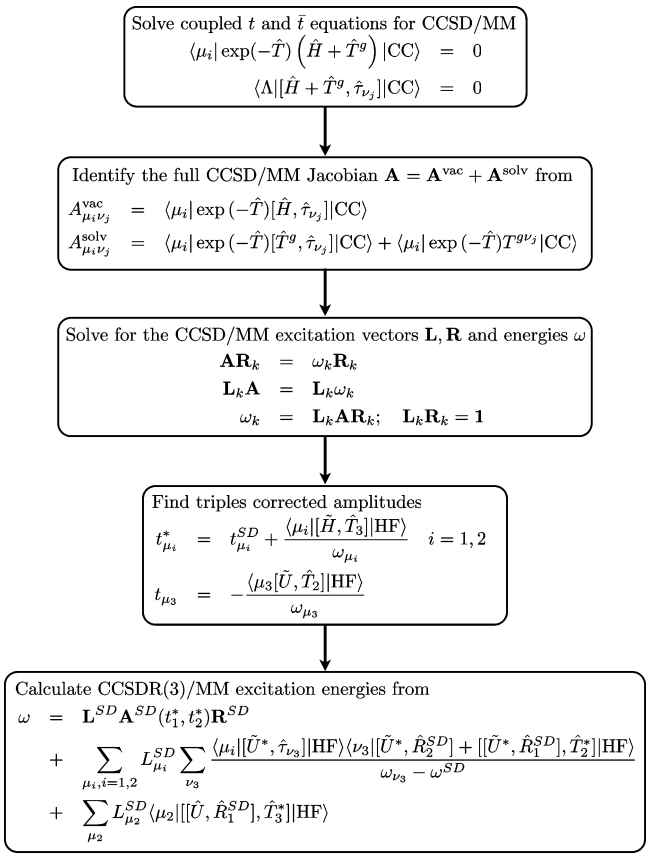


Figure 1. Overview of the proposed CCSDR(3)/MM model. See the text for definitions.

triples corrections, and we herein propose a CCSDR(3)/MM model which explicitly incorporates triples corrections in a noniterative fashion for the solute only, while the solvent−solute interaction is essentially described at the CCSD/MM level. Thus, we use the CCSD/MM vectors, in place of the vacuum CCSD vectors, as well as the triples-corrected operators in eq 7 and obtain a highly flexible description of the excitation energies of a molecule in a solvent. The proposed CCSDR(3)/MM model is illustrated in Figure 1.

3. Computational Details

The theory outlined in the previous section is here illustrated using various molecular systems: acrolein in aqueous solution, formamide in aqueous solution, NMA in aqueous solution, and liquid water. The effects of introducing a water solvent are incorporated by using molecular mechanics. The strategy behind the current QM/MM implementation is as follows: First, we determine partial charges and polarizabilites to be placed in the MM region. The latter are the key ingredient for determining induced dipoles allowing for a description of polarization between the two subsystems. Using these parameters a force field is constructed and subsequently applied in a molecular dynamics (MD) simulation using suitable settings for the macroscopic parameters (T, V, etc.), thus effectively simulating the effects of a nonzero temperature. From this simulation a set of 120 uncorrelated configurations, sufficient to obtain converged excitation energies,[21] is extracted and stored independently. Next we perform 120 independent QM/MM calculations until certain convergence criteria are satisfied; e.g., in the present implementation the induced dipoles are converged. Finally, we perform a statistical analysis of the 120 different excitation energies and oscillator strengths in order to estimate a number for the final valence excitation energy.

However, a straightforward averaging procedure over the different excitation energies ordered with respect to energy is problematic. Indeed, the excitations often become heavily mixed, as discussed in the next section. Thus, often it is very difficult to estimate a single vertical transition energy using a simple averaging technique. However, we may still construct a spectrum of the solvated sample, and to do so we introduce a broadening of the stick spectrum. We do this by assigning an explicit broadening using for each state of each configuration a Gaussian function of a finite width[23] (for simplicity fixed to 0.1 eV for all states in all molecules) designed such that the integral of each Gaussian gives the oscillator strength. We note that this identification is somewhat arbitrary and may be altered at our convenience if better resemblance to the experimental spectrum is strived for. The final spectrum will thus consist of a superposition of these distributions appropriately scaled by the inverse of the number of configurations to describe the averaged spectrum. Thus, the integral of the superposed Gaussians will give the averaged oscillator strength if one well-separated state is studied. This procedure has the advantage of giving a well-defined spectrum also in the case of mixed states, where the calculation of an average oscillator strength is problematic. The absorption strengths reported in some figures are the oscillator strength distributions as defined above. Finally, we compare the position of the band maximum to experiment, and if the experimental spectrum is readily available, we include a (scaled) version for ease of comparison. Note that we do not aim for exact agreement between the theoretical and experimental spectra, since we do not incorporate effects like vibronic coupling known to potentially change the appearance of the spectrum.

A basis set investigation[21] revealed that the basis set aug-cc-pVDZ[27] is sufficient for the description of the excitation energies of interest in acrolein. Similarly, for the NMA molecule we also apply the aug-cc-pVDZ basis, while we further augment diffuse functions (d-aug-cc-pVDZ) for the description of formamide. Finally, for the smaller water molecule the large d-aug-cc-pVTZ basis set is used.[22] We aim solely at improving the N-electron space, while the choice of basis set is kept fixed. All calculations use frozen 1s orbitals for the heavy atoms. Furthermore, a spherical cutoff radius of 12 Å (corresponding to the nearest 230−240 water molecules) is applied in all the QM/MM calculations. The geometry of the QM molecule(s) were all found using B3LYP/aug-cc-pVTZ accounting for the bulk solvent effects by using the familiar polarizable continuum model (PCM). The partial charges and polarizabilities were also determined at the B3LYP/aug-cc-pVTZ level of theory using the CHELPG procedure[28] and the LOPROP approach,[29] respectively. The Lennard-Jones parameters were taken from refs 30 and 31. The applied force fields were SPCpol for acrolein, formamide, and NMA, while TIP3P was used for liquid water. Further details on the MD simulations may be found

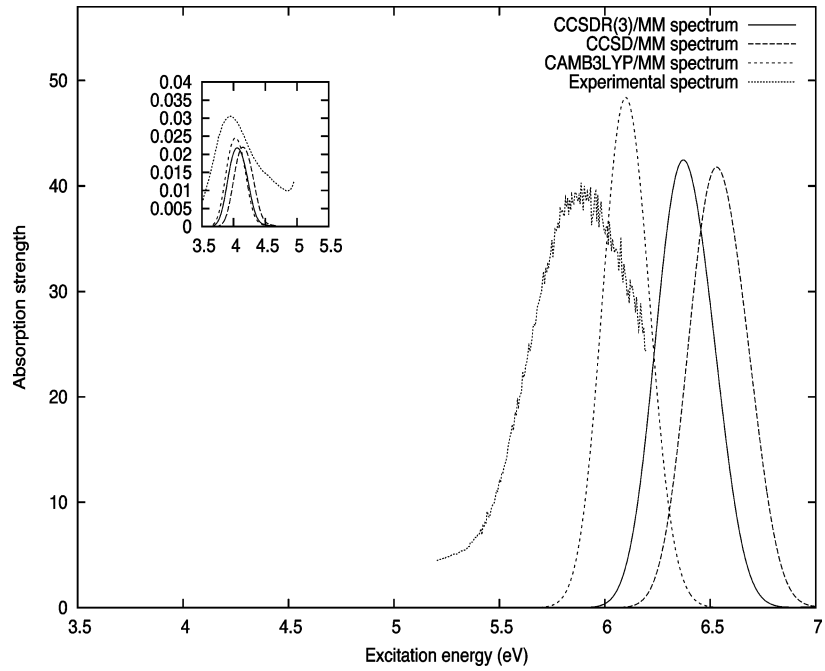


Figure 2. Acrolein spectrum n → π^* and $\pi \rightarrow \pi^*$ calculated at the CCSD/MM, CCSDR(3)/MM, and CAM-B3LYP/MM level of theory. For completion, an inset of the spectrum on a different scale is included. Also included is a scaled version of the experimental spectrum. See the text for details.

in ref 21. All vacuum calculations presented in the following are based on B3LYP/aug-cc-pVTZ-optimized geometries, as was also the case in the previous study of acrolein.[21]

The MD simulations have been performed using the MOLSIM program package.[32] The distributed polarizabilities were determined using the MOLCAS program.[33] The used QM/MM scheme was described in refs 34 and 35 and implemented in the DALTON quantum chemistry package.[36] The MIDASCPP program[37] has been used for preparing the input and the final statistical analysis of the data. Finally, the Gaussian 03 package[38] has been used for both finding the partial charges as well as optimizing the singlet excited states in order to estimate zero-point vibrational corrections to the energy.

4. Results

4.1. Acrolein in Aqueous Solution. As a first application of the CCSDR(3)/MM model, we make a thorough investigation of the two lowest-lying excitation energies of the acrolein molecule. In ref 21 it was found that the CAM-B3LYP functional compared well with findings using CCSD, with the obvious benefit of being much less computationally demanding, thus allowing for full QM calculations with as much as 12 water molecules in the QM part of the system. It was shown that this was indeed necessary in order to converge the $\pi \rightarrow \pi^*$ excitation energy, suggesting that this transition is dominated by nonelectrostatic contributions. In that paper, there is still small discrepancies between the converged CAM-B3LYP MD QM/MM(SPCpol) 12 (H_2O)QM and the experimental findings; these are the discrepancies we target by including triples. However, carrying out calculations on such large QM regions is not attractive, and the idea would thus be to estimate the contribution from a large QM solvent shell from CAM-B3LYP calculations. This will be clarified in the following.

Before this we note that it is not straightforward to compare theoretical and experimental values. First of all, in experiments a transition between electronic states is characterized by a relatively broad peak, making it difficult to directly deduce a vertical excitation energy; one has to assume this corresponds to the point of maximum absorption. As noted in ref 21, it is problematic to determine this maximum for the $\pi \rightarrow \pi^*$ transition in acrolein uniquely. As is clear from the experimental spectrum in Figure 2, the *point* of maximum absorbance is better described as an *interval* of maximum of absorbance with a length of approximately 0.5 eV, hence representing a substantial uncertainty considering the high level of theory used in this work. Second, effects due to intramolecular vibrations and the vibrational structure of the band are not included in the theoretical predictions.

With these words of caution we proceed by reporting the n → π^* excitations in Table 1. We note that this excitation is already fairly well described at the CCSD level. In Table 1 we also report the shifts referenced to acrolein in vacuum as well as the experimental and calculated values from ref 21.

It is seen that triples effects serve to lower the absolute values by about 0.1 eV, while the solvent shift is essentially unaltered. The shift is in good agreement with experiment, while the absolute value still is in disagreement by approximately 0.1 eV. Considering that this discrepancy is also present in the case of the vacuum transition, it is most probably not due to an insufficient solvent description but is related to the vibrational structure of the band not included in these calculations. This point is emphasized when considering the previously mentioned CAM-B3LYP calcula-

Table 1. Overview of the Reported Excitation Energies for the n → π* and π → π* Excitation Energies (in eV) of Acrolein in Vacuum and Water Solution and Corresponding Solvent Shifts[a]

ref	method	$^{vac}E_{ex}^{n\rightarrow\pi^*}$	$^{liq}E_{ex}^{n\rightarrow\pi^*}$	$\Delta E_{ex}^{n\rightarrow\pi^*}$	$^{vac}E_{ex}^{\pi\rightarrow\pi^*}$	$^{liq}E_{ex}^{\pi\rightarrow\pi^*}$	$\Delta E_{ex}^{\pi\rightarrow\pi^*}$
Aidas et al.[21]	CAM-B3LYP MD QM/MM(SPCpol) 0 (H_2O)QM	3.78	4.14	0.36	6.41	6.13	−0.28
	CAM-B3LYP/MM(SPCpol) 2 (H_2O)QM		4.04	0.26		6.07	−0.34
	CAM-B3LYP/MM(SPCpol) 12 (H_2O)QM		4.04	0.26		5.95	−0.46
present work	CCSD MD QM/MM(SPCpol) 0 (H_2O)QM	3.91	4.16	0.25	6.87	6.54	−0.33
	CCSD/MM(SPCpol) 2 (H_2O)QM		4.15	0.24		6.51	−0.36
	CCSDR(3)/MM(SPCpol) 0 (H_2O)QM	3.81	4.07	0.26	6.73	6.38	−0.35
	CCSDR(3)/MM(SPCpol) 2 (H_2O)QM		4.06	0.25		6.35	−0.38
	CCSDR(3)/MM(SPCpol) 0 (H2O)QM + E_{nonel} (CAM-B3LYP-SPCpol)		4.08	0.27		6.22	−0.51
expt[21]		3.69	3.94	0.25	6.41	5.90	−0.52

[a] Also included are the values calculated in a related previous study as well as experimental results.

tions on acrolein,[21] which are also included in Table 1. Explicitly, it is noted that the n → π* has no significant contribution from water molecules when these are explicitly included in the QM part.

Finally, we note that a recent CASSCF//CASPT2 investigation[39] gave vacuum and solvated n → π* transition energies of 3.77 and 3.96 eV, respectively, thus providing a shift in good agreement with those presented Table 1.

We now turn our attention to the π → π* excitation in Table 1. Comparing the CCSDR(3) and CCSD calculated values, we see a larger effect of triples excitations for this state, resulting in a lowering of approximately 0.2 eV for the absolute excitation energy. As for the n → π* transition the impact of triples is close to being canceled when considering solvent shifts rather than absolute values of the excitation energies.

Focusing on the shifts we see that, compared to the experimental values, the CCSDR(3) values are still somewhat off, suggesting that the remaining error is not due to triples effects but rather due to the nonelectrostatic nature of the interaction between this excited state and the solvent.

We estimate the size of this effect by taking the difference between two CAM-B3LYP calculated excitation energies. We subtract the excitation energy found with a QM region including only acrolein itself from a large-scale QM calculation (12 water molecules treated quantum mechanically; see ref 21 for a justification for the size of the QM system). This contribution will be added to the excitation energy found using CCSDR(3) including also only acrolein in the QM region. The results are included in Table 1 and this extra contribution is labeled E_{nonel}(CAM-B3LYP-SPCpol). For completion, we have included the n → π* nonelectrostatically corrected excitation energies. We see that the shifted values with the nonelectrostatic correction are almost identical to the experimental values. Though this final agreement between theory and experiment is comforting, one should not overemphasize this. Especially, we note that since the CCSD/MM and CCSDR(3)/MM solvent shifts only differ by approximately 0.02 eV we could have performed a similar estimate for the CCSD/MM excitation energy shifts. Large-scale CAM-B3LYP calculations turned out in this case to be a very efficient and economical way to estimate the nonelectrostatic effects not included in the (purely electrostatic) interaction between the QM and MM subsystems.

Now changing our focus to the absolute excitation energies in the gas phase, we see that they are still alluding us by almost 0.3 eV as compared with experiment. This deviation might suggest that the remaining contribution is not connected with the lack of electron correlation but rather has to do with geometry effects (all QM calculations are performed on B3LYP-optimized structures as in ref 21) and the lack of vibrational structure of the band (including zero-point energy contributions).

Table 2. Lowest-Lying Singlet Excitation Energies (in eV) for Water in Vacuum and the Liquid Phase as Well as the Corresponding Solvatochromic Shift Calculated in a Hierarchy of CC Methods as Well as with DFT

ref	method	$^{vac}E_{ex}^{\tilde{A}}$	$^{liq}E_{ex}^{\tilde{A}}$	$\Delta E_{ex}^{\tilde{A}}$
present work	B3LYP/MM	6.90	7.43	0.53
	CAM-B3LYP/MM	7.13	7.72	0.59
	CC2/MM	7.25	7.86	0.61
	CCSD/MM	7.62	8.25	0.63
	CCSDR(3)/MM	7.61	8.25	0.64
expt[42]		7.4	8.2	0.8

Consequently, we have estimated the zero point vibrational contribution (ZPVC) to the lowest-lying singlet excited state for acrolein in a vacuum. Using CIS/6-311++G(d,p) to optimize the excited state structure followed by an evaluation of the vibrational frequencies, we estimate this contribution to be −0.11 eV, which combined with the CCSDR(3) vacuum excitation energy (3.81 eV − 0.11 eV = 3.70 eV) is in excellent agreement with experiment (3.69 eV). The vibrational structure of the second excited state is noteworthy more complex, as also noted elsewhere.[40] However, using the vibrational frequencies available in ref 40 at the CASSCF/cc-pVTZ level of theory as well as performing an analogous calculation for the ground state (see ref 40 for details on the active space) we are able to estimate a ZPVC of approximately −0.14 eV, which combined with the CCSDR(3) calculated excitation energy (6.73 eV − 0.14 eV = 6.59 eV) is in satisfactory agreement with experiment (6.41 eV) when recalling the very diffuse nature of the experimental band. The fact that CAM-B3LYP is in such good agreement *without* inclusion of these effects is perhaps a little worrisome but nevertheless very remarkable.

On the basis of oscillator strengths from CCSD/MM calculations, we have also constructed CCSDR(3)/MM spectra for acrolein in aqueous solution. In Figure 2 the CCSD/MM as well as CCSDR(3)/MM spectra are given. We recall that the spectra are calculated by representing each

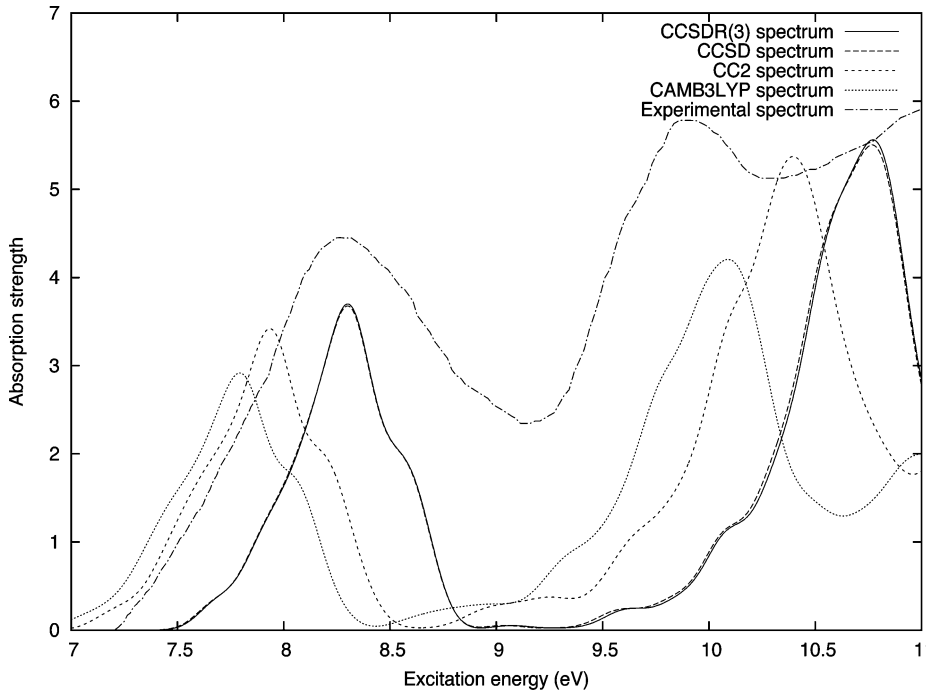


Figure 3. Water spectrum, with the lowest-lying excitations in water averaged over 120 configurations calculated at CC2/MM, CCSD/MM, CCSDR(3)/MM, and CAM-B3LYP/MM levels of theory. Also included is a scaled version of the experimental spectrum.

Table 3. Overview of the Reported Excitation Energies for the n → π^* and $\pi \rightarrow \pi^*$ Excitation Energies (in eV) of Formamide in Vacuum and Water Solution and Corresponding Solvent Shifts[a]

method	$^{vac}E_{ex}^{n\rightarrow\pi^*}$	$^{liq}E_{ex}^{n\rightarrow\pi^*}$	$^{vac}E_{ex}^{\pi\rightarrow\pi^*}$	$^{liq}E_{ex}^{\pi\rightarrow\pi^*}$
CAM-B3LYP/MM(SPCpol)	5.59	5.97	7.69	7. 45[b]
CCSD/MM(SPCpol)	5.70	6.15	7.58	7.16
CCSDR(3)/MM(SPCpol)	5.69	6.13	7.69	7.09
Expt[43,48]	5.8	5.2−5.9	7.4	>6.5

[a] Also included are the experimental results. [b] Assignment based on spectrum.

state for each structure by a Gaussian whose integral is proportional to the oscillator strength.

It is evident that the inclusion of triples red-shifts the band. Due to symmetry reasons the n → π^* transitions in carbonyl compounds are typically weak and it is therefore of no surprise that the oscillator strength (not given explicitly here) of the n → π^* excitation is approximately 3 orders of magnitude smaller than for the $\pi \rightarrow \pi^*$ transition demonstrated by the inset in Figure 2. Finally, we have also included the CAM-B3LYP/MM spectrum in Figure 2.

4.2. Liquid Water. We have also investigated the electronic spectrum of water with the CCSDR(3)/MM model. We will focus only on the lowest-lying excitation—in the literature often labeled $\tilde{A}$—since this, as discussed in ref 23, is well-separated (by approximately 1 eV) from the remaining ones. In Table 2 we show the lowest-lying excitation energy calculated for liquid water modeled explicitly by one water molecule being treated quantum mechanically while the remaining ones are treated classically. We see that for this system DFT gives far too low excitation energies stemming from the partly Rydberg nature of the electronic transitions in water. This is not surprising, since the approximate exchange functionals used in DFT contain spurious electronic self-interaction terms which previously have been noted to give too low Rydberg excitation energies.[41] On the contrary, full inclusion of doubles in the CCSD model markedly increases the excitation energies while the final inclusion of triples has only fine-tuning effects of approximately 0.03 eV. This conclusion is even more apparent in Figure 3, where we have superimposed four theoretical spectra at the CC2/MM, CCSD/MM, CCSDR(3)/MM, and CAM-B3LYP/MM level of theory, respectively, as well as a recreation of the experimental spectrum.[42] Going from CC2 to CCSD clearly results in a blue-shifting of the bands, while inclusion of triples effects leaves an almost identical spectrum compared to CCSD. This is as expected considering the electronic structure of water. We also note that the CAM-B3LYP spectrum is red-shifted as compared to the CC2 spectrum, thus emphasizing the underestimation of the excitation energies of water when described by DFT.

In experiment, the location of the two lowest-lying excitations is approximately 8.2 eV in good agreement with the calculated $\tilde{A}$ excitation energy.

4.3. Formamide in Aqueous Solution. Drawing on the conclusions from the acrolein investigation, we apply the CCSDR(3)/MM model on formamide. We especially compare CCSDR(3)/MM and CAM-B3LYP results and the resulting spectra. In most amides the spectrum is characterized by, besides a multitude of Rydberg transitions, a very intense $\pi \rightarrow \pi^*$ transition and a very weak n → π^* transition. Upon introduction of a polar solvent (e.g., water) these two valence transitions are red- and blue-shifted, respectively. Consequently, the weak n → π^* transition is hidden below the much stronger $\pi \rightarrow \pi^*$ band, naturally complicating a thorough characterization of this vertical transition. Therefore, we here perform an extensive

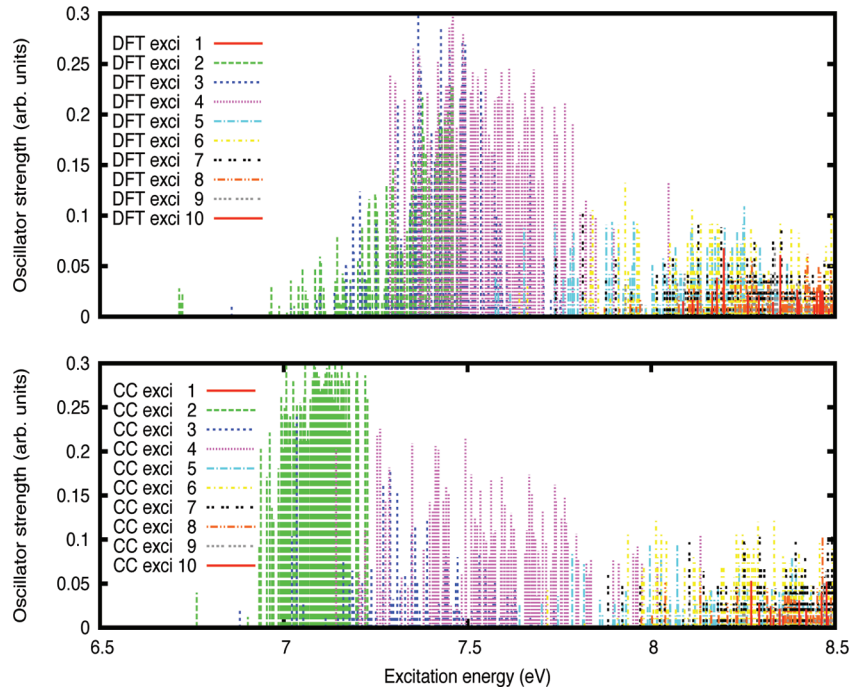


Figure 4. Distribution of the 10 lowest-lying excitation energies in 120 different configurations of formamide solvated in water described using CCSDR(3)/MM and CAM-B3LYP/MM, respectively.

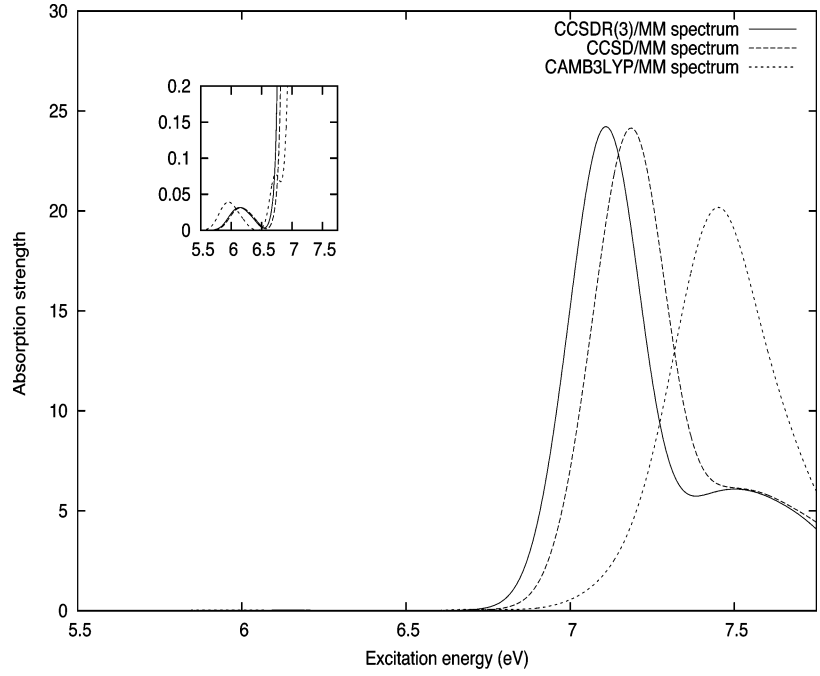


Figure 5. CCSD/MM, CCSDR(3)/MM, and CAM-B3LYP/MM spectra of aqueous formamide with insets in different scale to illustrate the weak n → π^* transition.

analysis of the degree of mixing between the excited states energies and take measures accordingly when estimating the valence excitation energies. In doing so we also hope that the technical challenges involved in simulating realistic spectra of solvated samples are illuminated.

In Table 3 we have included the calculated excitation energies of the lowest-lying valence excitations in formamide. For the n → π^* excitation there is a discrepancy between the calculated vertical excitation energy and the experimental energy of maximum absorption of approximately 0.1 eV. Given that inclusion of triples effects takes us further away from the experimental value, we conclude that the remaining effects are not due to dynamical correlation. Similarly, for the $\pi \rightarrow \pi^*$ transition the CC and DFT models are in disagreement with experiment by approximately 0.3 eV, where we again see that inclusion of triples does not improve the description of the excitation energy as compared to experiment. On the basis of the acrolein study it seems likely that the remaining disagreement between theory and experiment is due primarily to ZPVC effects. Actual calculations, implying yet another excited state optimization, is beyond the scope of this paper.

Table 4. Overview of the Reported Excitation Energies for the n → π* and π → π* Excitation Energies (in eV) of NMA in Vacuum and Water Solution and Corresponding Solvent Shifts[a]

method	$^{vac}E_{ex}^{n\rightarrow\pi^*}$	$^{liq}E_{ex}^{n\rightarrow\pi^*}$	$^{vac}E_{ex}^{\pi\rightarrow\pi^*}$	$^{liq}E_{ex}^{\pi\rightarrow\pi^*}$
CAM-B3LYP/MM(SPCpol)	5.75	6.15	6.41	7. 17[c]
CCSD/MM(SPCpol)	5.88	6.30	6.43	7. 05[c]
CCSDR(3)/MM(SPCpol)	5.84	6.27	6.29	6. 96[c]
expt[46,49]	N/A	5. 54[b]	6.81	6.67

[a] Also included are the experimental results. [b] Estimated using solvent difference techniques. [c] Assignment based on spectrum in Figure 7.

Instead we focus on the analysis of the CCSDR(3)/MM results. In Figure 4 we show the distribution of the excitation energies in the 120 different configurations. In this figure we do not directly observe the weak n → π* transition since (1) it is well-separated (lower) in energy from the rest and (2) the oscillator strengths are several orders of magnitude smaller than the other valence excitation. Therefore, it is straightforward to find a point of maximum absorption by averaging the lowest-lying excitation energy in all the configurations, thus obtaining a value of 6.13 eV. However, the second lowest excitation energy does not always correspond to a π → π* transition (based on the size of the oscillator strength). That being said, the mixing only occurs for a relatively few number of configurations and it is still possible to estimate a π → π* excitation energy by an explicit averaging over the second-lowest excitation energy obtaining a value of 7.09 eV. Also depicted in Figure 4 is a similar analysis using CAM-B3LYP as opposed to CCSDR(3) for the description of the QM region. We observe that for CAM-B3LYP it is essentially impossible to deduce a pure π → π* transition, as the excitation energies are heavily mixed, preventing any meaningful averaging. Finally, we may construct a full spectrum; indeed, one might argue that it is the only viable option in cases where the Rydberg transitions and valence transitions are so heavily mixed. The final spectra for CCSD/MM, CCSDR(3)/MM, and CAM-B3LYP/MM are shown in Figure 5. We note that a peak corresponding to the n → π* is located at approximately 6.1 eV for CCSDR(3)/MM and at 6.0 eV for CAM-B3LYP/MM. The π → π* excitation energy is peaked at approximately 7.1 eV for CCSDR(3)/MM and 7.5 eV for CAM-B3LYP/MM. Also, in the CAM-B3LYP/MM spectrum we observe a small kink on the low-energy side of the strong π → π* band. In the experimental spectrum[43] a band in the tail of the π → π* band is observed from 5.2 to 5.9 eV and assigned to the n → π* transition. This is somewhat lower than our predicted values. For the π → π* excitation energy no experimental details are available, except that it is located above 6.5 eV, as predicted by both models. Finally, we note the application of other methods for calculating vertical excitation energies and shifts. In particular, a macroscopic continuum CASSCF//CASPT2 investigation[44] gave n → π* and π → π* vertical transition energies of 5.54 and 6.95 eV, respectively, while a Monte Carlo INDO-CIS calculation[45] provided extrapolated solvent shifts of 0.2 eV and −0.1, respectively.

4.4. NMA in Aqueous Solution. As a further test of CCSDR(3)/MM, we investigate the spectrum of a slightly more complicated amide: *N*-methylacetamide (NMA). In Table 4 we have included the calculated vacuum excitation energies of the lowest-lying valence excitations. The n → π* excitation energy is not available in experiment while the π → π* excitation energy peaks at 6.81 eV.

In Figure 6 we demonstrate the distribution of the lowest-lying excitations in NMA for all 120 configurations using both CCSDR(3) and CAM-B3LYP. Contrary to the case of formamide, we see that not even CCSDR(3) offers a way to easily estimate a π → π* excitation energy. We also observe that only the CC calculations reveal a degree of mixing

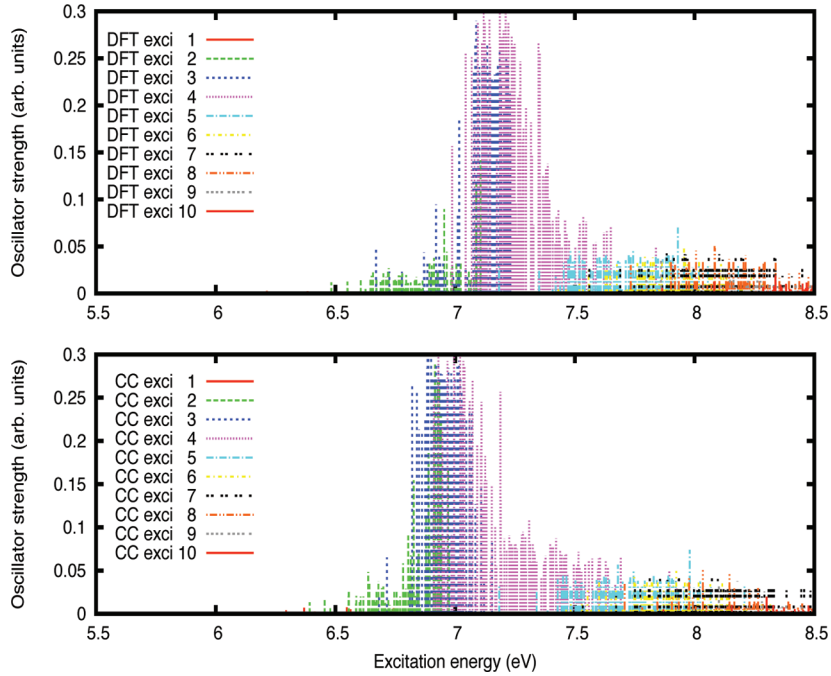


Figure 6. Distribution of the 10 lowest-lying excitation energies in 120 different configurations of NMA solvated in water described using CCSDR(3)/MM and CAM-B3LYP/MM, respectively.

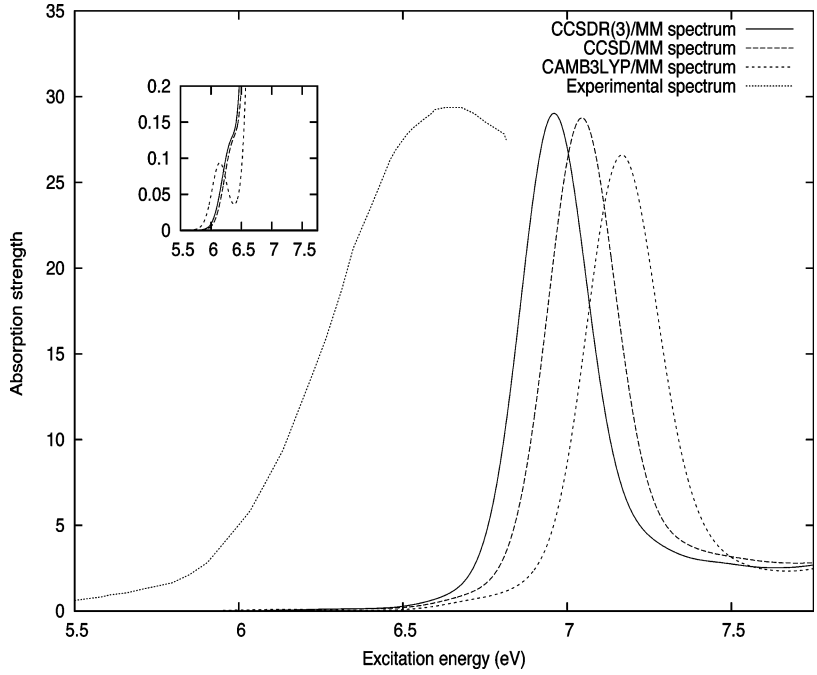


Figure 7. CCSD/MM, CCSDR(3)/MM, and CAM-B3LYP spectra of aqueous NMA with insets in different scale to illustrate the weak n → π^* transition. Also included is a scaled version of the experimental spectrum.

between the weak n → π^* and the higher-lying transitions. At a first inspection, the picture provided by DFT (a well-separated n → π^* transition) may seem more comforting. However, as hinted to in the previous section, in experiment it is extremely difficult to estimate the n → π^* excitation energy of amides, as it is essentially hidden (oscillator strength of the order 0.0025 was reported in fine agreement with the values presented here) under the much stronger π → π^* band. The complicated photoabsorption of NMA is not apparent from the DFT calculations but is conveyed nicely in the CC calculations. Finally, we must be cautious especially when analyzing the DFT spectrum since, as mentioned in the water investigation, DFT has problems describing excitations that are Rydberg in nature.

We have constructed the final spectrum of NMA in aqueous solution shown in Figure 7 and from this we may estimate approximate band maxima. In experiment a band with a maximum at 6.7 eV is assigned to the π → π^* transition. This is in agreement with the CCSDR(3)/MM spectrum in Figure 7, which has a strong bandwidth at approximately 6.9 eV. When no triples effects are included (e.g., the CCSD/MM model) we observe a slight overestimation of this transition as compared to experiment while the CAM-B3LYP/MM calculations seem to overestimate the position of this band even further. In experiment, a value for the n → π^* excitation energy has been estimated (5.5 eV) using nonpolar → polar solvent difference techniques, but it was also noted that these methods usually provide values that are too low.[46] From Figure 7 we see a very weak shoulder at around 6.3 eV. For the spectrum calculated using CAM-B3LYP/MM we see, as mentioned above, that the n → π^* is more separated from the strong π → π^* transition as compared to the CC calculation. For completeness we have included a recreation of the experimental spectrum.[46]

In the literature other calculations of vertical excitation energies and shifts have been reported. As for the formamide molecule we note that a macroscopic continuum CASSCF//CASPT2 calculation[44] gave a n → π^* excitation energy of 5.56 eV and a π → π^* energy of 6.60 eV. Finally, Monte Carlo INDO-CIS extrapolated solvent shifts have been reported[47] for the considered states of approximately 0.22 and −0.15 eV, respectively, both shifts being somewhat lower than the values reported in Table 4.

5. Conclusions

We have introduced the CCSDR(3)/MM model, a noniterative method to incorporate triples effects in a QM/MM calculation. We have tested it on four model systems: *s-trans*-acrolein, formamide, and NMA in aqueous solution and liquid water. For the former system we perform a thorough investigation of the two lowest-lying valence singlet excitation energies. For the n → π^* we obtain perfect agreement with experiment for the solvent shift. The vacuum energy comes close to experiment only when including ZPVC. The π → π^* proves a little more troublesome as it contains a large nonelectrostatic contribution requiring a large QM region. This was estimated using CAM-B3LYP, ultimately yielding good agreement between theory and experiment for the solvent shift. Ultimately, the effects of triples in acrolein on the lowest singlet excitation energies correspond to approximately −0.1 and −0.15 eV. The investigation of two amides (formamides and NMA) shows us that it is not straightforward to include the effects of dynamics in a simple averaging procedure, as the individual excitation energies tend to mix in the different configurations. While CCSDR(3)/MM proved a solution for formamide, difficulties arise in

the NMA spectrum, allowing essentially only for a meaningful comparison between band maxima and the experimental values.

For liquid water we find good agreement between experiment and theory for the lowest studied singlet excitation energy. The spectrum calculated using CCSDR(3)/MM energies is almost identical to the one calculated using CCSD/MM, implying negligible triples effects for this system.

Acknowledgment. This work has been supported by the Lundbeck Foundation and DCSC (Danish Center for Scientific Computing). O.C. acknowledges support from the Danish national research foundation, the Lundbeck Foundation, and EUROHORCs through a EURYI award. J.K. thanks the Villum Kann Rasmussen Foundation and the Danish Natural Science Research Council/The Danish Councils for Independent Research for financial support.

References

(1) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(2) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

(3) Singh, U.; Kollman, P. *J. Comput. Chem.* **1986**, *7*, 718.

(4) Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *J. Chem. Phys.* **2003**, *118*, 1620.

(5) Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. The (Hyper)Polarizabilities of Liquid Water Modelled Using Coupled Cluster/Molecular Mechanics Response Theory Methods. In *Atoms, Molecules and Clusters in Electric Fields. Theoretical Approaches to the Calculation of Electric Polarizability*; Maroulis, G., Ed.; Imperial College Press: London, 2006.

(6) Kowalski, K.; Valiev, M. *J. Phys. Chem. A* **2006**, *110*, 13106.

(7) Fang, P.; Valiev, M.; Kowalski, K. *Chem. Phys. Lett.* **2008**, *458*, 205.

(8) Epifanovsky, E.; Kowalski, K.; Fan, P.-D.; Valiev, M.; Matsika, S.; Krylov, A. I. *J. Phys. Chem. A* **2008**, *112*, 9983.

(9) Hirata, S.; Yagi, K. *Chem. Phys. Lett.* **2008**, *464*, 123.

(10) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.

(11) Christiansen, O.; Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1995**, *103*, 7429–7441.

(12) Christiansen, O.; Koch, H.; Jørgensen, P.; Helgaker, T.; Olsen, J. *Chem. Phys. Lett.* **1996**, *256*, 185.

(13) Christiansen, O.; Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1996**, *105*, 1451.

(14) Sauer, S. P. A.; Schreiber, M.; Silva-Junior, M. R.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 555.

(15) Manohar, P. U.; Stanton, J. F.; Krylov, A. I. *J. Chem. Phys.* **2009**, *131*, 114112.

(16) Fukuda, R.; Hayaki, S.; Nakatsuji, H. *J. Chem. Phys.* **2009**, *131*, 174303.

(17) Watts, J. D.; Bartlett, R. J. *Spectrochim. Acta, Part A* **1999**, *55*, 495.

(18) Shiozaki, T.; Hirao, K.; Hirata, S. *J. Chem. Phys.* **2007**, *126*, 244106.

(19) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2003**, *120*, 1715.

(20) Fujimoto, K.; Hayashi, S.; Hasegawa, J.; Nakatsuji, H. *J. Chem. Theory Comput.* **2007**, *3*, 605.

(21) Aidas, K.; Møgelhøj, A.; Nilsson, E. K.; Johnson, M. S.; Mikkelsen, K. V.; Christiansen, O.; Söderhjelm, P.; Kongsted, J. *J. Chem. Phys.* **2008**, *128*, 194503.

(22) Christiansen, O.; Nymand, T.; Mikkelsen, K. V. *J. Chem. Phys.* **2000**, *113*, 8101.

(23) Osted, A.; Kongsted, J.; Mikkelsen, K. V.; Aastrand, P. O.; Christiansen, O. *J. Chem. Phys.* **2006**, *124*, 124503.

(24) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*, 1st ed.; Wiley: New York, 2000.

(25) Christiansen, O.; Koch, H.; Halkier, A.; Jørgensen, P.; Helgaker, T.; de Meras, A. S. *J. Chem. Phys.* **1996**, *105*, 6921.

(26) Christiansen, O.; Mikkelsen, K. V. *J. Chem. Phys.* **1999**, *110*, 1365.

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

(28) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(29) Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494.

(30) Georg, H. C.; Coutinho, K.; Canuto, S. *J. Chem. Phys.* **2005**, *123*, 124307.

(31) Xie, W.; Pu, J.; MacKerell, A.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1878.

(32) Linse, P. *MOLSIM, an integrated MD/MC/BD simulation program belonging to the MOLSIM package in C++*, 2004.

(33) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.

(34) Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *Mol. Phys.* **2002**, *100*, 1813.

(35) Nielsen, C. B.; Christiansen, O.; Mikkelsen, K. V.; Kongsted, J. *J. Chem. Phys.* **2007**, *126*, 154112.

(36) *DALTON, a Molecular Electronic Structure Program, Release 2.0*, 2005; see http://www.kjemi.uio.no/software/dalton/dalton.html,

(37) *Midascpp, Molecular Interactions, Dynamics and Simulation Chemistry Program Package in C++*; http://www.chem.au.dk/∼midas (accessed Jan 6, 2010).

(38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.;

Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*, Gaussian Inc.: Wallingford, CT, 2004.

(39) Losa, A. M.; Galvan, I. F.; Aguilar, M. A.; Martin, M. E. *J. Phys. Chem. B* **2007**, *111*, 9864.

(40) Bokareva, O. S.; Bataev, V. A.; Popyshev, V. I.; Godunov, I. A. *Spectr. Acta Part A* **2009**, *73*, 654.

(41) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1999**, *314*, 152.

(42) Heller, J. M.; Hamm, R. N.; Birkhoff, R. D.; Painter, L. R. *J. Chem. Phys.* **1974**, *60*, 3483.

(43) Petersen, C.; Dahl, N. H.; Jensen, S. K.; Poulsen, J. A.; Thøgersen, J.; Keiding, S. R. *J. Phys. Chem. A* **2008**, *112*, 3339.

(44) Besley, N. A.; Hirst, J. D. *J. Phys. Chem. A* **1998**, *102*, 10791.

(45) Rocha, W. R.; Martins, V. M.; Coutinho, K.; Canuto, S. *Theor. Chem. Acc.* **2002**, *108*, 31.

(46) Nielsen, E. B.; Schellman, J. A. *J. Phys. Chem.* **1967**, *71*, 2297.

(47) Rocha, W. R.; Almeida, K. J. D.; Coutinho, K.; Canuto, S. *Chem. Phys. Lett.* **2001**, *345*, 171.

(48) Gingell, J. M.; Mason, N. J.; Zhao, H.; Walker, I. C.; Siggel, M. R. F. *Chem. Phys.* **1997**, *220*, 191.

(49) Kaya, K.; Nagakura, S. *Theor. Chim. Acta.* **1967**, *7*, 124.

CT900641W

JCTC Journal of Chemical Theory and Computation

A Coarse-Grained Model Based on Morse Potential for Water and *n*-Alkanes

See-Wing Chiu,*,† H. Larry Scott,‡ and Eric Jakobsson§

Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, Illinois 61801, Department of Biological, Chemical and Physical Sciences, Illinois Institute of Technology, Illinois 60616, Department of Molecular and Integrative Physiology, Beckman Institute for Advanced Science and Technology, Department of Biochemistry, Center for Biophysics and Computational Biology, University of Illinois, Urbana, Illinois 61801

Received September 8, 2009

Abstract: In order to extend the time and distance scales of molecular dynamics simulations, it is essential to create accurate coarse-grained force fields, in which each particle contains several atoms. Coarse-grained force fields that utilize the Lennard-Jones potential form for pairwise nonbonded interactions have been shown to suffer from serious inaccuracy, notably with respect to describing the behavior of water. In this paper, we describe a coarse-grained force field for water, in which each particle contains four water molecules, based on the Morse potential form. By molecular dynamics simulations, we show that our force field closely replicates important water properties. We also describe a Morse potential force field for alkanes and a simulation method for alkanes in which individual particles may have variable size, providing flexibility in constructing complex molecules comprised partly or solely of alkane groups. We find that, in addition to being more accurate, the Morse potential also provides the ability to take larger time steps than the Lennard-Jones, because the short distance repulsion potential profile is less steep. We suggest that the Morse potential form should be considered as an alternative for the Lennard-Jones form for coarse-grained molecular dynamics simulations.

1. Introduction

There has been a growing interest in applying coarse-grained (CG) superatom models for molecular dynamics (MD) simulations for a variety of polymers. These CG models consist of superatoms (or beads) that represent groups of atoms, or even several molecules. Coarse graining makes it possible to study larger systems at remarkably longer time scales with acceptable levels of detail. An overall protocol for moving between atomically detailed models and coarse-grained models for polymers by coarse graining and reverse coarse graining was described by Müller-Plathe.[1] The central method is to use a potential of mean force derived from atomistic simulations to parametrize the coarse-grained model, although the parameters can also be tuned by reference to thermodynamic properties. The coarse-grained simulations serve as an "express highway" to an equilibrated state, which is then reverse coarse-grained to atomic detail to reveal a detailed structure of the equilibrated state. A detailed example of the Müller-Plathe approach to coarse graining is given in Reithe et al.[2] An example of the full cycle of coarse-graining followed by simulation (in this case, Monte Carlo), in turn followed by reverse mapping, is given in Spyriouni et al.[3] Wilson et al.[4] provided a review of multiscale simulations as applied to liquid crystals, including macromolecular liquid crystals. A particular feature of simulations of liquid crystals is the ability to use hybrid models in which one component of a molecule (the mesogenic component that induces the ordering characteristic

* Corresponding author e-mail: s-chiu@illinois.edu.

† Beckman Institute for Advanced Science and Technology, University of Illinois.

‡ Illinois Institute of Technology.

§ Center for Biophysics and Computational Biology, University of Illinois.

10.1021/ct900475p
Published on Web 02/17/2010

of liquid crystals) is modeled as being completely rigid, while the rest of the molecule that determines the liquid character is modeled as flexible.[4] Prampolini[5] utilized a variant of this approach in which the properties of the mesogenic component are parametrized by ab initio calculations.

The MARTINI CG force field field[6,7] is especially designed for the simulation of lipids, surfactants, and proteins. This force field is parametrized by extensive calibration of the chemical building blocks of the CG force field against thermodynamical data. Because the MARTINI CG force field is especially designed for a variety of biomolecular simulations (our group's primary application interest), and because it is currently under active development, we take it as a primary reference point for our work in this area. However, we note the existence of coarse-grained simulations of amphiphilic molecules in aqueous environments using other force fields, such as Loison et at,[8] Shelley et al.,[9] and Markvoort et al.[10] A polarizable CG water model has been proposed by Ha-Duong et al.[11]

Another coarse-grained force field of biomolecular interest is the M3B force field,[12] designed to simulate maltooligosaccharides and their interaction with water. A theoretical point of interest is that the nonbonded non-Coulombic interactions in M3B are done with a Morse potential,[13] rather than with a potential of the Lennard-Jones form. In the M3B force field, each water molecule is represented by a single particle interacting with others by a Morse potential. The Morse potential was originally developed as a mathematically efficient way to describe chemical bond formation and dissociation.[13] In ref 12, the applicability of the Morse potential is effectively extended to the making and breaking of hydrogen bonds. Liew and Masuhiro[14] suggest that the application of a Morse-like potential may be an effective general strategy to improve the ability of coarse-grained simulations to properly represent phase changes and coexistence. This point speaks directly to a major problem in the MARTINI force field, which is based on using a Lennard-Jones (LJ) 12−6 interaction potential.[15] The major problem is the fact that its water model freezes at physiological temperatures.[16]

In this paper, we report on the application of a Morse potential to particle−particle interactions in a water model at the MARTINI level of coarse graining, which is four water molecules per particle. We will compare the behavior of this new coarse grained water model to the MARTINI model and to a coarse grained water model due to Shelly et al.,[9] where they have applied a softer 6−4 potential to their CG water model which represents a group of three water molecules. We also develop CG *n*-alkane models based on using the Morse potential.[13] We will refer to the CG water model of Shelley et al.[9] as SSRBK, the acronym of the last names of the authors of the reference. Similarly, we will refer to our models as CSJ.

2. Morse potential

The Morse potential $V_M(r)$, has the form

$$V_M(r) = \varepsilon[e^{\alpha(1-\frac{r}{R_0})} - 2e^{1/2\alpha(1-\frac{r}{R_0})}] \quad (1)$$

In eq 1, R_0 is the distance of the minimum energy ε and α is a parameter that measures the curvature of the potential around R_0. The smaller the value of α, the softer is the potential. In the condensed phase, the density of the system is mainly affected by the value of R_0, the cohesive energy by ε, and the compressibility by α.[12]

In the present work, we use eq 1 as the pair interaction potential function for the CG *n*-alkanes and water (group of four water molecules). The adjustable parameters ε and R_0 are parametrized to fit the experimental densities and heats of vaporization of liquid *n*-alkanes and water via CG simulations. Computational detail is to be described in the Computational Method section. The selection of the values of α for the CG models for *n*-alkanes and water is based on the agreement of the simulated vapor−liquid interfacial tensions with the experimental data.

3. Computational Method

The recent 43A1-S3 atomistically detailed force field for hydrocarbons and lipids[17] was used to carry out MD simulations of the bulk phase for each of the *n*-alkanes from *n*-butane to *n*-heptadecane. The MD trajectories of these hydrocarbon liquids were taken from our previous work.[17] Intramolecular data from these atomistically detailed simulations were mapped according to the CG models as described below and were used in the course of CG parametrization.

GROMACS 4.0.4 modeling software[18] was used for MD simulations performed in this work. User-custom lookup tables for the Morse potential (eq 1) were prepared according to the format as described in the manual of the software,[19] and the LJ parameters in the interaction parameter file for CG atoms were all set to 1.0.

For CG-MD simulations, a time step (Δt) size of 40 fs was generally used unless otherwise specified explicitly. We note that the use of such too large an integration time step in CG-MD may produce energy sinks and induce errors.[16] Because of the softer nature of the repulsive component of the Morse potential function (eq 1), we found that use of a time step (Δt) of 40 fs for CG-MD simulations is permissible in this work. Energy fluctuations, $\Delta E = \langle [E - \langle E \rangle]^2 \rangle^{1/2}$, as a function of Δt were performed to evalulate how large Δt could be used in this work.[16] Details of the test are presented in the Supporting Information. A cutoff of 1.6 nm was used for eq 1 without using switching or shifting the function and for the pair list updating. It should be noted that, when the energy lookup table is used, GROMACS 4.0.4 uses the neighbor search cutoff as the real cutoff for the interaction potential. Neighbor searching is usually done with a larger radius than the cutoff specified for the potential to accommodate for the size of charge groups and diffusion between neighbor list updates. Our CG models carry no partial charges. Each CG interaction site was treated as one charge group. Nonbonding interactions involving the first neighbors in *n*-alkane chains were excluded in all CG-MD calculations, and no bond length constraint such as SHAKE was applied. The nonbonding pair list was updated every 5 time steps. Temperature boundary conditions were set using the Nose-Hoover algorithm.[20] A temperature coupling constant of 0.5 ps was used for $\Delta t = 40$ fs. It was reduced to 0.2 ps when

a smaller time step of 10 fs was used. All simulations were performed at 298 K unless otherwise explicitly specified.

Bulk calculations were performed out for both water and liquid alkanes. For water, calculations were performed on systems of two sizes, 400 water molecules and 3200 water molecules. There was no significant size-dependent effect. For liquid alkane calculations, there were 400 molecules in the butane simulation and 200 molecules in the simulations for the larger alkane molecules.

Calculations were also carried out for systems where the liquid coexisted with a vapor. In those calculations, there were 3200 water molecules, 3200 butane molecules, and 1600 alkane molecules, respectively.

3.1. Enthalpy of Vaporization. For MD simulations of bulk phases of CG water and *n*-alkanes, NPT boundary conditions with isotropic pressure coupling were applied. Pressure boundary conditions were set using the Parrinello−Rahman pressure coupling method.[21] The pressure coupling constant was set to 5 ps for $\Delta t = 0.04$ or 2 ps for $\Delta t = 0.01$ ps. For runs to determine molecular volume and heat of vaporization, each system was simulated for a time length of 20 ns. The enthalpy of vaporization ΔH_{vap} of a liquid was calculated from two simulations for its liquid phase and its gas phase, as previously described:[22]

$$\Delta H_{vap} = V_{Intra}(g) - [V_{Inter}(l) + V_{Intra}(l)] + RT = -V_{Inter}(l) + RT \quad (2a)$$

where V_{Intra} is the intramolecular potential energy per mole of molecules calculated for both the gas (g) and the liquid (l) states and V_{Inter} is the intermolecular energy per mole of molecules. The use of eq 2a is based on the assumption of ideal gas behavior and the assumption that the sum of kinetic and vibrational energies is equal for the gas and liquid states. When eq 2a is applied to calculate ΔH_{vap} for CG water, the term $V_{Inter}(l)$, which represents the inter-CG-water energy per mole of CG water, does not include the intermolecular interaction $V_{S,n}$ among the *n subunits* (water molecules) of each CG site. Theoretically, a correction including $V_{S,n}$ to ΔH_{vap} is required in order to compare directly with the experimental ΔH_{vap}. Thus, for an *n*:1 mapped CG water,

$$\Delta H_{vap} = -\frac{V_{Inter}(l) + V_{S,n}}{n} + RT \quad (2b)$$

One may approximate the binding energy of a water cluster formed by *n* molecules of water in the gas phase, which can be calculated quantum mechanically, as $-V_{S,n}$. The ab initio binding energies of the water trimer (61.9 kJ/mol)[23] and tetramer (128.5 kJ/mol)[24] in their lowest energy configurations have been calculated at a high theoretical level. We use these values as the values for the internal energy in eq 2b in analyzing our computational results.

3.2. Free Energy of Solvation. For simulating a liquid in equilibrium with its vapor, a slab (about 7 nm in thickness) of the equilibrated bulk phase of a CG species under consideration was placed in the center of a simulation box with a vacuum slab thickness of about 3.5 nm on each side. The application of periodic boundary conditions creates a system of alternating liquid and vapor layers, with the liquid layer being 7-nm-thick and the vapor layer being 7-nm-thick. The CG-MD simulation was performed in the NVT ensemble at 298 K. During the simulation, a small number of molecules enter the vapor phase (there are no molecules in the vapor phase at the beginning of the simulation), and ultimately a dynamic steady state is reached in which molecules are evaporating and condensing at approximately the same rate. For water and for the shorter alkanes, the simulations were done for 10 μs, with the last 5 μs trajectories being used for data analysis. For long-chain *n*-alkanes (*n*-tetradecane, *n*-pentadecane, etc.), a time length of 20 μs was used in order to obtain meaningful statistics, because of the smaller number of molecules evaporating. The Gibbs free energy of solvation of a gaseous solute in its own liquid can be calculated from the equilibrium densities of the particles in its vapor phase (ρ_v) and its bulk phase (ρ_l) according to Ben-Naim and Marcus,[25,26]

$$\Delta G_S = -RT \ln\left(\frac{\rho_l}{\rho_v}\right) = -RT\ln\left(\frac{c_l RT}{p_v}\right) \quad (3)$$

where c_l is the molar concentration of the liquid and p_v is the vapor pressure. Since vapor−liquid systems were simulated under the NVT condition, the calculated ΔG_S values are in fact Helmholtz free energies ΔF_S from which $\Delta G_S = \Delta F_S + \Delta(PV)$ can be obtained. Note that the $\Delta(PV)$ correction term applies to the condensed phase before and after solvation of the solute molecules according to Ben-Naim and Marcus' definition of the solvation process[25] and is usually negligible.[19]

Thermodynamic integration (TI) procedure[27] was also carried out under NPT conditions to calculate ΔG_S for alkanes. The free energy of solvation was evaluated over a path that mutated all CG sites in a single CG solute molecule into noninteracting ones via a coupling parameter λ in such a way that $\lambda = 0$ describes the noninteracting phantom site and $\lambda = 1$ describes the fully interacting CG site via the following λ-dependent Morse potential function:

$$V_M(r;\lambda) = \lambda\varepsilon[e^{\lambda\alpha(1-\frac{r}{\lambda R_0})} - 2e^{1/2\lambda\alpha(1-\frac{r}{\lambda R_0})}] = \lambda\varepsilon[e^{\alpha(\lambda-\frac{r}{R_0})} - 2e^{1/2\alpha(\lambda-\frac{r}{R_0})}] \quad (4)$$

Thus, eq 4 is equivalent to eq 1 when $\lambda = 1$, and $V_M(r;\lambda)$ is zero when $\lambda = 0$. In the process of gradual mutation from $\lambda = 0$ to $\lambda = 1$, the internal bonded and nonbonded interactions were kept at their full value so that the integrity of the solute molecule was maintained. The process can be viewed as gradual coupling of a single solute molecule in a vacuum and a box of pure solvent. (Note that, in this terminology, the molecule that is being created by "computational alchemy" is the "solute", and the rest of the molecules in the system comprise the "solvent".) In TI, the integral,

$$\Delta G_S = \int_0^1 \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda d\lambda \quad (5)$$

is used to evaluate ΔG_S, where *H* is the classical Hamiltonian which depends on configuration variables and momenta. Since there is no λ dependence of the internal energy terms of the solute molecule, it does not contribute directly to

$\partial H(p,q;\lambda)/\partial\lambda$. In addition, there is no mass changed in our coupling scheme; there is also no kinetic-energy contribution to $\partial H(p,q;\lambda)/\partial\lambda$. Hence, eq 5 simply becomes

$$\Delta G_S = \int_0^1 \left\langle \frac{\partial V_M(r;\lambda)}{\partial\lambda} \right\rangle_\lambda d\lambda \quad (6)$$

An analytical derivative of eq 4 was used to evaluate $\langle \partial V_M(r;\lambda)/\partial\lambda \rangle$.

Simulations were started at $\lambda = 1$ from a well equilibrated box (ca. $4 \times 4 \times 4$ nm^3) of about 200 to 400 CG alkanes, depending on the size of the alkane. An alkane (solute) molecule fully interacting with the rest of its own kind of molecules (solvent) was randomly chosen from the initial configuration. The λ interval value was set at 0.05. However, in the region of rapidly changing $\langle \partial V_M(r;\lambda)/\partial\lambda \rangle$, it was reduced to 0.025. Subsequent simulation at the next λ interval value was performed using the last configuration from the simulation at a previous λ value. Each simulation consisted of a 5 ns equilibration period followed by a 15 ns production run. An integration time step of 10 fs was used. The $\langle \partial V_M(r;\lambda)/\partial\lambda \rangle$ value was evaluated for every ps. The integration of eq 6 was performed by the trapezoidal rule.

Since the current GROMACS (version 4.0.4) is not capable of performing the TI procedure based on the Morse potential, the above TI method was carried out manually by using the energy-group-lookup-table option available in the MD software. For each λ value, a potential energy lookup table for the interaction between the target solute and the solvent (eq 4) was set up according to the specification of GROMACS 4.0.4. For solvent−solvent and solute−solute nonbonded interactions, a separate interaction lookup table (eq 1) was made. A separate code was written to calculate $\langle \partial V_M(r;\lambda)/\partial\lambda \rangle$ with the configurations extracted from the simulations trajectories.

3.3. Interfacial Tension. The vapor−liquid interfacial tension γ, a measure of the free energy cost associated with the formation of the interface, was computed from the ensemble average normal P_{ZZ} and lateral P_{XX}/P_{YY} pressure components according to

$$\gamma = \frac{1}{2} L_Z \left(\left\langle P_{ZZ} - \frac{P_{XX} + P_{YY}}{2} \right\rangle \right) \quad (7)$$

The factor 1/2 accounts for the two interfaces present in the chosen setup. In general, a μs of CG simulation was necessary to achieve reliable statistics for the computation of γ. The implementation of eq 7 is contained in a standard Gromacs utility, which we used.

3.4. Coarse-Grained Modeling. *Water Tetramer.* The CG water atom (Figure 1), namely W4, is a single bead which represents a group of four water molecules similar to the MARTINI water model.[6] For comparison, Figure 1 shows the corresponding interaction potentials for the MARTINI model and the SSRBK[9] model. (Note for the SSRBK model that the SSRBK CG particle contains only three waters rather than four, so its potential form is not strictly comparable with the other two.) The CG particles interact through $V_M(r)$, eq 1. The target data for determining the parameters of eq 1 for W4 are the experimental liquid water density and the

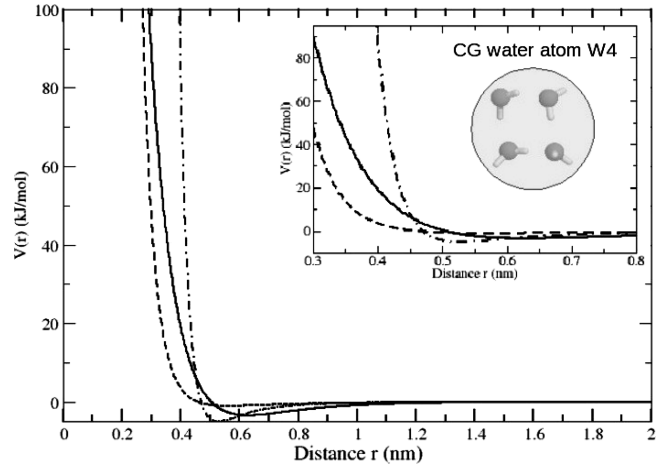


Figure 1. Coarse grained water model and its interaction potential. The CG W4 atom represents a cluster of four water molecules. Solid line, CSJ water (Morse potential); dashed line, SSRBK water[9] (LJ6−4 potential); dot-dashed line, MARTINI water[6] (LJ12−6 potential). It should be noted that the CSJ and MARTINI models are for four-water clusters, while the SSRBK model is for a three-water cluster.

Table 1. Morse Parameters for the CG Water (W4) and the Interaction Sites of 3-Site (C3T, C3M) and 4-Site (C4T, C4M) Mapped Alkanes

interaction site	α	ε (kJ/mol)	R_0 (nm)	atomic mass
W4	7	3.4	0.629	72.062
C3T	12	2.94	0.527	43.089
C3M	12	2.94	0.527	42.081
C4T	12	4.0	0.563	57.116
C4M	12	4.0	0.563	56.108

interfacial tension of air−water or water vapor−water. For a series of MD runs with different α values, we optimized ε and R_0 to have calculated ρ and γ in agreement with the experimental values. The final values of α, ε, and R_0 for W4 so obtained are listed in Table 1. We did not use the experimental ΔH_{vap} of water as a target in the course of parametrization for W4, but very good values for ΔH_{vap} were obtained, as mentioned later in the Results and Discussion section. There, we will examine more closely this and other similar CG water models.

n-Alkanes. Alkanes with Multiples of 3 or 4 Carbon Atoms. Two CG types of particles for modeling alkanes are employed, 3-site (C3) and 4-site (C4) mapped CG particles. The former is suitable for alkanes with a multiple of three carbon atoms, and the latter is for those with a multiple of four carbon atoms. For example, *n*-hexane is modeled as a linear chain of two C3 CG atoms (Figure 2a), while octane can be represented by a linear chain of two C4 CG atoms, as shown in Figure 2a.

For the 3-site (4-site) CG model, the CG alkanes were constructed from terminal C3T (C4T) and middle C3M (C4M) chain units. These CG units interact through eq 1. In terms of the nonbonding (Morse potential) parameters, the C3T and C4T are not differentiated from C3M and C4M, respectively. However, they are regarded as separate atom types because their masses are slightly different.

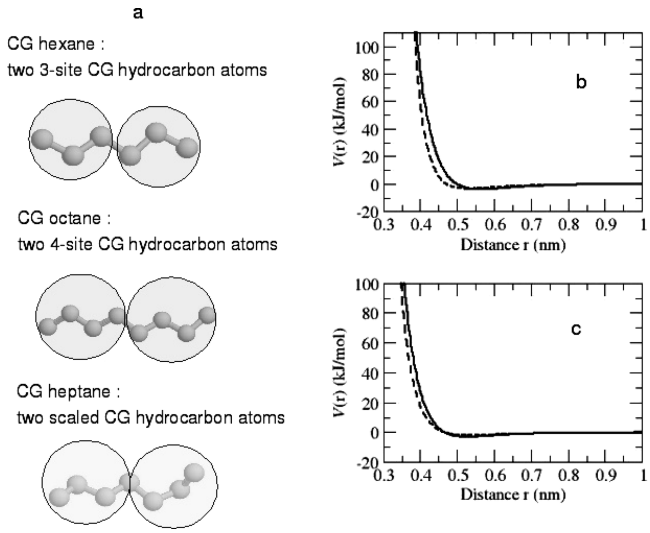


Figure 2. Coarse grained alkane models and their interaction potentials. (a) Hexane is represented by two 3-site CG hydrocarbon atoms. Octane is represented by two 4-site CG hydrocarbon atoms. Heptane is represented by two scaled CG hydrocarbon atoms. (b) Pair interaction potentials for 4:1 mapped CG site. Solid line: Morse potential (this work). Dashed line: LJ 12−6 potential (MARTINI).[6] (c) Pair interaction potentials for 3:1 mapped CG site. Solid line: Morse potential (this work). Dashed line: LJ 9−6 potential (of Nielsen et al.[32]).

Masses for the CG types C3T, C3M, C4T, and C4M are listed in Table 1.

Intramolecularly, the CG alkanes interact through harmonic bond and angular bend potentials,

$$V_{\text{bond}} = \frac{K_{\text{b}}}{2}(r_{\text{b}} - r_{\text{b0}})^2 \tag{8}$$

$$V_{\text{angle}} = \frac{K_{\text{a}}}{2}(\cos\theta - \cos\theta_0)^2 \tag{9}$$

All first neighbors were excluded from nonbonding interaction. The target CG bond lengths (r_{b0}) were calculated from atomistic simulation of dodecane as performed from our previous work.[17] These trajectories were first mapped according to the CG model used. Starting from one end of the alkane chain, n ($n = 3$ for 3-site mapping; $n = 4$ for 4-site mapping) consecutive carbon atoms were counted and their center of mass as the location of the first CG site was calculated. The second CG site was similarly determined by counting further n carbon atoms. Actual masses were used for center of mass calculations. The average distances of two consecutive CG sites were respectively computed from these 3:1 and 4:1 mapped MD trajectories. These are the target bond lengths (r_{b0}) for the CG models and are listed in Table 2. The force field parameters $K_a = 25$ kJ/mol and $\theta_0 = 180°$ (eq 9) were taken from the CG MARTINI force field[6] without further refinement. They are listed in Table 3. It should be noted that the MARTINI n-alkanes and our force field as derived in the main body of this paper do not have torsion potentials. The consequence of not taking torsion potential into account is a uniform dihedral distribution for the CG model, whereas a corresponding atomistic model shows a bias toward lower torsion angles.[29] In further simulations done since the main body of simulations reported in this paper, we find that including torsion angle restraints improves dramatically the agreement of the dihedral angle distributions but has little effect on the thermodynamic properties. Details of the parametrization method and results are presented in the Supporting Information. This extended parameter set includes torsion potential as well as distinction between terminal and nonterminal CG sites in terms of nonbinding interactions. All its nonbonding, bonding, bending, and torsion parameters are parametrized against experimental and atomistic data. In the course of developing this parameter set, we found that setting θ_0 equal to 180° instead of the mean of pseudo-bond angles mapped from atomistic data results in much better correspondence between CG and atomistic bending angle distributions.

Table 2. CG Standard Bond Length r_{b0} and Force Constant K_b Parameters

bond type	r_{b0} (nm)	K_b (kJ mol^{-1} nm^{-2})
C3T, C3M−C3T, C3M	0.36	5000
C4T, C4M−C4T, C4M	0.45	5000

Table 3. CG Standard Bond Angle θ_0 and Force Constant K_a Parameters for 3-Site and 4-Site CG Models

angle type	θ_0 (degree)	K_a (kJ mol^{-1})
C3T, C3M−C3M−C3T, C3M	180	25
C4T, C4M−C4M−C4T, C4M	180	25

CG simulations of liquids hexane and octane were initially performed using the MARTINI bond force constant K_b for alkanes (1250 kJ mol^{-1} nm^{-2}). For different values of α, the parameters ε and R_0 were adjusted so that the simulated ρ and ΔH_{vap} of hexane and octane were in agreement with the experimental values.[30] Their vapor−liquid interfacial tensions were also computed with these initially parametrized sets of $\{\alpha,\varepsilon,R_0\}$. We found that $\{\alpha,\varepsilon,R_0\}$ with $\alpha = 12$ yielded γ values close to the experimental data.[31] In addition, K_b was tuned by matching the radial distribution functions (RDF) for CG hexane and octane to those obtained from mapped atomically detailed MD data. The K_b value so determined, 5000 kJ mol^{-1} nm^{-2}, is good for both 3-site and 4-site CG alkanes. We now have $\alpha = 12$ and $K_b = 5000$ kJ mol^{-1} nm^{-2} set for all CG alkanes.

With the parametrized values of $\alpha = 12$ and $K_b = 5000$ kJ mol^{-1} nm^{-2} and other bonding parameters as listed in Tables 2 and 3, the ε and R_0 were refined by carrying out CG simulations of pentadecane and hexadecane. The target data were their experimental ρ and ΔH_{vap}. The parameters were finally modified slightly so that the CG simulated ρ and ΔH_{vap} were in good agreement with the experiment for both the short (hexane, octane) and long (pentadecane, hexadecane) chain alkanes. Table 1 lists the final values of Morse parameters (α,ε,R_0) for both 3-site and 4-site mapped CG alkanes. Figure 2b shows the nonbonded interaction potentials for the 4-carbon alkane CG interactions in the CSJ

and MARTINI force fields. Figure 2c shows the nonbonded interaction potentials for the 3-carbon alkane CG interactions in the CSJ (this paper) and the NLSK[32] force fields.

Scaling for Alkanes. For chain length with non-multiple-of-three or multiple-of-four carbon atoms, a combination of 3-site and 4-site CG atoms can be introduced. For example, *n*-heptane can be represented by one C3T and one C4T atom. However, this option leads to a nonuniform representation of the chain. In what follows, we introduce a scaling method for the Morse parameters to avoid such unwanted choice.

The number of carbon atoms in an alkane chain, N_C, is divided by three, and the quotient, an integer, gives the number of CG sites, N_{CG}. Each site is then scaled in mass and in center of mass placement by $W_S = N_C/N_{CG}$. In other words, the *extra* carbon atoms (N_C modulo 3) are spread uniformly over the CG chain; i.e., they are evenly shared among the N_{CG} sites. To map the atomistic alkane chain onto a scaled CG representation, W_S carbons are counted starting from one end of the atomistic chain. The center of mass of these atoms determines the location of the first CG site. A reduced mass is given to fractional atoms in the center of mass calculation. The second CG site is then assigned by counting the next W_S carbons and determining their center of mass. If the last counting ends at a fractional site of the atomistic chain, its remainder is used for the current counting. The CG representation of *n*-undecane, for example, has (quotient of 11/3) 3 CG sites. The number of extra carbon atoms is (N_C modulo 3) 2, which are equally shared among the three CG sites. Each CG site has scaling weight of $W_S = N_C/N_{CG} = 11/3$. The first CG site is the center of mass of the first three carbon atoms and 2/3 of the fourth one. The center of mass of the remainder (1/3) of the fourth carbon, its following three carbon atoms, and 1/3 of the eighth carbon atom is then the location of the second CG site. To assign the location of the third CG site, the center of mass of the remainder (2/3) of the eighth carbon atom and the last three carbon atoms is computed.

To obtain the target CG bond lengths (r_{b0}) for *n*-heptane, *n*-decane, *n*-undecane, *n*-tridecane, and *n*-heptadecane, the aforementioned scaling method was used to map the atomically detailed trajectories of these *n*-alkanes from our previous work[17] to their CG representations according to their scaling weights. The time average of the distances between two consecutive CG sites was computed from the atomically detailed simulations for each scaled CG representation, and the result was assigned as the r_{b0} of those CG alkanes whose CG sites have the same scaling weight. Thus, CG *n*-heptane and *n*-tetradecane have the same r_{b0} value. The same bond and bond angle force constants as used in the 3-site/4-site models were applied, and θ_0 was set to 180°. All these parametric values are listed in Table 4.

To calculate R_0 for the scaled CG sites, we assumed that the increase in R_0 from 3-site CG ($W_S = 3$) to 4-site CG ($W_S = 4$) is linear with the increased scaling weight. Hence, the R_0 for a CG site with scaling weight W_s can be simply calculated from

$$R_0(s) = 0.527 + \Delta R_0(W_s - 3) \tag{10}$$

Table 4. Scaled CG Standard Bond Length r_{b0}, Bond Angle θ_0, and Their Force Constants K_b and K_a

alkane	W_S	r_{b0} (nm)	K_b (kJ mol^{-1} nm^{-2})	θ_0 (deg)	K_a (kJ mol^{-1})
heptane, tetradecane	$3^1/_2$	0.400	5000	180	25
decane	$3^1/_3$	0.390	5000	180	25
undecane	$3^2/_3$	0.413	5000	180	25
tridecane	$3^1/_4$	0.371	5000	180	25
heptadecane	$3^2/_5$	0.375	5000	180	25

where the first term is the R_0 value for the 3-site CG model (Table 1), ΔR_0 (0.036 nm) is the difference of the R_0 values between the 4-site and 3-site CG atoms, and the numerical value 3 in the last term is the scaling weight of the 3-site CG atom.

To obtain the ε values of the scaled CG sites, we employed the following combination rule: We compute the geometrical mean value of the parameters ε of two CG sites with known W_S values of w_1 and w_2, respectively. The result is assigned as the ε value for a CG site with a W_S value which is the arithmetic mean of w_1 and w_2:

$$\varepsilon\left(\frac{w_1 + w_2}{2}\right) = (\varepsilon(w_1)\,\varepsilon(w_2))^{1/2} \tag{11}$$

For CG *n*-heptane and *n*-tetradecane with a scaling weight of $3^1/_2$, their $\varepsilon(3^1/_2)$ value was calculated from the already known $\varepsilon(3)$ and $\varepsilon(4)$ values of the 3-site and 4-site CG atoms, respectively (Table 1), according to eq 11. The scaling weight for CG *n*-tridecane is $3^1/_4$. Hence, its $\varepsilon(3^1/_4)$ value was calculated from $\varepsilon(3)$ and $\varepsilon(3\ 1/2)$ using eq 11. The scaling weight for *n*-decane is $3^1/_3$. Its $\varepsilon(3^1/_3)$ value cannot be evaluated in the same way as previously mentioned. Instead, it was approximated by successively computing the value of $\{\varepsilon(w_i)\}$ according to eq 12

$$\varepsilon(w_{i+2}) = (\varepsilon(w_i)\,\varepsilon(w_{i+1}))^{1/2}, w_{i+2} = \frac{w_{i+1} + w_i}{2} \tag{12}$$

For $w_1 = 3\ 1/2$ and $w_2 = 3\ 1/4$, the fisrt two terms of the $\{\varepsilon(w_i)\}$ sequence are the already known $\varepsilon(w_1) = \varepsilon(3\ 1/2)$ and $\varepsilon(w_2) = \varepsilon(3^1/_4)$ values as computed for *n*-heptane and *n*-tridecane, respectively. The third term, $\varepsilon(w_3) = \varepsilon(3^3/_8)$, was then calculated using eq 12. The fourth term $\varepsilon(w_4) = \varepsilon(3^5/_{16})$ was then in turn computed with the known values of $\varepsilon(w_2)$ and $\varepsilon(w_3)$. This procedure was repeated until the term $\varepsilon(w_i)$ with $w_i \approx 3^1/_3$ was obtained. The scaled $\varepsilon(3^2/_3)$value for CG *n*-undecane was computed in the same way using eq 12 with the known $\varepsilon(w_1) = \varepsilon(4)$ and $\varepsilon(w_2) = \varepsilon(3^1/_2)$ values as the first and second terms of $\{\varepsilon(w_i)\}$, respectively. The iterative computation was proceeded until the term $\varepsilon(w_i)$ with $w_i \approx 3^2/_3$ was obtained.

In scaling $\varepsilon(3^2/_5)$ for CG *n*-heptadecane, we first calculated the series (eq 12) through the term $\varepsilon(3^3/_{16})$ starting with the known values of $\varepsilon(w_1) = \varepsilon(3)$ and $\varepsilon(w_2) = \varepsilon(3^1/_4)$. The $\varepsilon(3^{13}/_{32})$ value was then evaluated from $\varepsilon(3^3/_{16})$ and $\varepsilon(3^5/_8)$ according to eq 11. The latter had already been obtained in the course of calculating $\varepsilon(3^2/_3)$. Since $3^{13}/_{32}$ is approximately equal to $3^2/_5$, we hence assigned the so calculated $\varepsilon(3^{13}/_{32})$ as the value of $\varepsilon(3^2/_5)$ for CG *n*-heptadecane.

Table 5. Morse Parameters for Scaled CG Sites[a]

alkane	CG type	W_S	ε (kJ/mol)	R_0 (nm)	α
heptane, tetradecane	CST, CSM	$3^{1/2}$	3.43	0.545	12
decane	CST, CSM	$3^{1}/_{3}$	3.26	0.539	12
undecane	CST, CSM	$3^{2}/_{3}$	3.61	0.551	12
tridecane	CST, CSM	$3^{1}/_{4}$	3.175	0.536	12
heptadecane	CST, CSM	$3^{2}/_{5}$	3.33	0.541	12

[a] Terminal sites are denoted as CST and non terminal sites are designated as CSM. Actual masses for CG sites are used according to their scaling weight W_S.

Table 6. Calculated Density ρ (g/cm^3), Self Diffusion Coefficient D (10^{-9} m^2/s), Heat of Vaporization ΔH_{vap} (kJ/mol), Free Energy of Solvation ΔG_S (kJ/mol), Surface Tension (water-vapor) γ (mN/m), and Isothermal Compressibility κ (10^{-5} bar^{-1}) of CG Water W4 at 298 K for CG Water Models[a]

Water model	ρ	D	ΔH_{vap}	ΔG_S	γ	κ
CSJ[b] W4	0.998	4.3	38.4	−28	71	17
CSJ[b] W4	*0.996*[f]	*4.7*[f]	*38.3*	*−28*[f]	*71*[f]	*26*[f]
MARTINI[c] W	1.005	1.6	30.2	−18	32	9
SSRBK[d] W	0.993	6.6	32.6	−19	71	15
Experiment[e]	0.998	2.3	44.0	−26.5	73	4.5

[a] CG-MD integration time step used was 10 fs. [b] This work. Equation 1 is the pair interaction potential. [c] Ref 6, CG-MD simulations were performed in this work. [d] SSRBK is the acronym of the last names of the authors of ref 9. CG-MD simulations were performed in this work. [e] Refs 30 and 31. The experimental ΔG_S was calculated from the vapor pressure p_v (ref 31) using eq 3. [f] Italic numbers were calculated from simulations performed using an integration time step of 40 fs.

All the scaled Morse parameters obtained in this section are listed in Table 5. They were applied for CG simulations without further parametrization.

4. Results and Discussion

4.1. CG Water. Table 6 lists the calculated density ρ, heat of vaporization ΔH_{vap}, free energy of solvation ΔG_S, and liquid−vapor interfacial tension γ of CSJ water. There, the listed numerical data for MARTINI[6] and SSRBK[9] CG water were also calculated for comparison. The simulation conditions used for these comparisons essentially replicated those used by the workers who developed those force fields except that in order to do the comparison we used a time step of 10 fs, a pairlist update frequency of 5 time steps, the Nose-Hoover temperature coupling method, the and Parrinello−Rahman algorithm for pressure coupling. Both CSJ (this work) and MARTINI[6] CG water models employ 4:1 mapping while the SSRBK[9] model has a single bead to represent three water molecules. The Martini water interacts via a LJ 12−6 potential. The pair potential for the SSRBK water is a softer LJ 6−4. As can be seen from Figure 1, the LJ 12−6 potential for Martini water is much harder than the other two in the short-range region. Using a large integration time step such as 40 fs with a cutoff of 1.2 nm as is done in the MARTINI water model induces errors[16] and produces energy sinks which cause a freezing effect. We found that using a smaller time step of 10 fs with the MARTINI force field does not remove the freezing artifact but delays its occurrence. We hypothesize

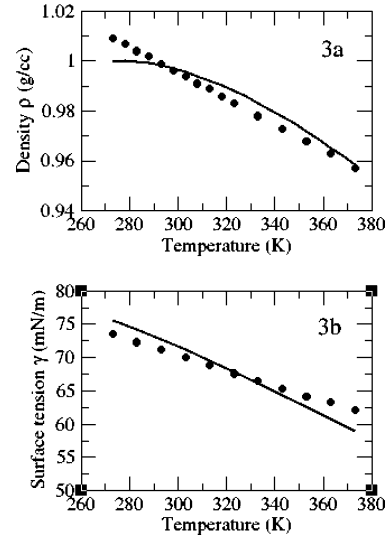


Figure 3. Density ρ and surface tension γ of CG water. (a) Temperature dependent density of CG water: CSJ water (black circle), experiment (solid line). (b) Temperature dependent surface tension of water: CSJ water (black circle), experiment (solid line).

that the fundamental problem lies in the use of the LJ 12−6 potential, which produces a narrow deep well as shown in Figure 1. The SSRBK water model (LJ 6−4) may alleviate this problem but provides only two free parameters, which reduces flexibility in tuning the potential to fit a variety of experimental data. There exists an advantage of using the Morse potential for targeting atomistic or experimental data, in that one can control more freely the interactions for $r \leq R_0$ and $r > R_0$ independently by using different α values for $r \leq R_0$ and $r > R_0$ whenever it is desired. Another 3-parameter candidate for pair interaction is the Buckingham potential[33] (BP) $V_{BP}(r) = b \exp(-r/\rho) - \mu/r^6$ where b, ρ, and μ are constants and r is the interatomic distance. There are several studies in which a Buckingham form (near repulsion is an exponential form) has been used in pairwise interaction potentials.[34−39] A concern in using such a form is that at very short distances, the attractive sixth power term overwhelms the repulsive exponential term and the potential becomes very large and negative, which is nonphysical. It would preclude the use of the Buckingham potential in the early stages of building macromolecular complexes, in which the systematic and automated relief of steric clashes is needed. The Morse potential (eq 1) does not diverge at very small distances and has a finite positive value at $r_{ij} = 0$. We do not find this a problem in our simulations, because with the chosen values of α, the repulsive potential is still large enough to prevent the coalescence of the interacting pair.

Under normal physiological temperatures, both the CSJ and the SSRBK water models are designed to have their intensive properties ρ and γ be consistent with those of water. Both CSJ and SSRBK water models have calculated γ values (71 nN/m at 298K) in good agreement with the measured value, 73 mN/m.[30] Figure 3 shows the temperature dependence of ρ and γ for CSJ water. Both ρ and γ of CSJ water

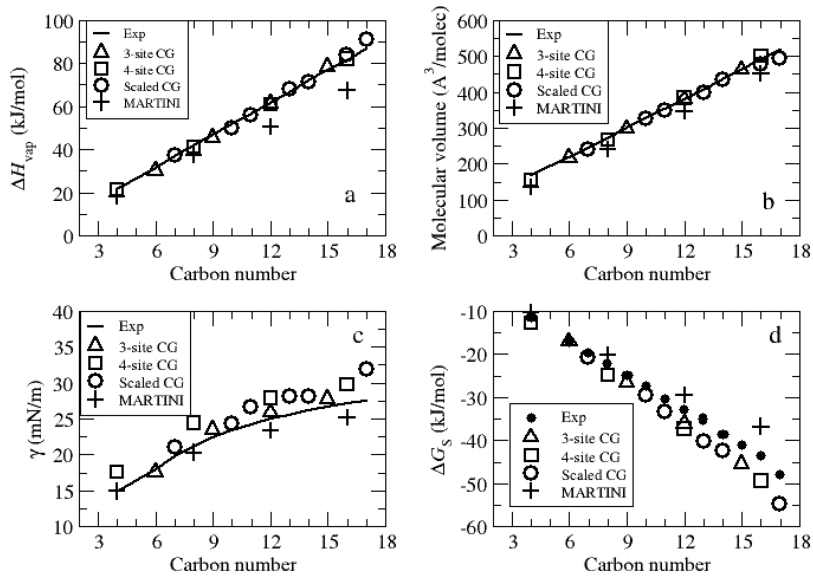


Figure 4. Physical and thermodynamics properties of alkanes at 298 K. (a) Enthalpies of vaporization ΔH_{vap}. (b) Molecular volumes of alkanes. (c) Vapor−liquid interfacial tension γ of alkanes. Data for butane were obtained at 273 K. (d) Gibbs solvation free energies ΔG_S of alkanes.

Table 7. Solvation Free Energy ΔG_S of *n*-Alkanes at 298 K

	ΔG_S (kJ/mol)		
n-alkane	TI (eq 5)	vapor−liquid (eq 3)	experiment[a]
butane	−12.8	−12.6	−11.5
hexane	−17.1	−17.4	−16.9
heptane	−20.8	−21.1	−19.7
octane	−24.9	−25.4	−22.3
nonane	−26.6	−27.3	−24.9
decane	−29.5	−30.1	−27.5
undecane	−33.5	−34.2	−30.5
dodecane	−36.2(−37.3)[b]	−37.2(−36.1)[b]	−33.0
tridecane	−40.3	−42.0	−35.0
tetradecane	−42.3	−43.0	−38.8
pentadecane	−45.5	−43.0	−41.1
hexadecane	−49.4	−50.9	−43.7
heptadecane	−54.9	−52.4	−48.1

[a] Ref 31. The experimental ΔG_S was calculated from the vapor pressure p_v using eq 3. [b] Numbers in parentheses are for the 4-site CG model.

decrease linearly as T increases. The CG water model does not however capture the nonlinear temperature dependence of the experimental data in the temperature range 273−373 K. As can be seen in Figure 3, although the general trends and overall ranges of the properties are similar to the experiment, the forms of the calculated density and surface tension are much more linear than the experimental curves. This is likely a consequence of the loss of some atomistic details that occurs in the coarse graining process. This shows limits of the accuracy which can be expected for coarse-grained calculations as soon as quantities involving derivatives are needed.

The Gibbs excess free energy ΔG_{ex} of water is equivalent to the negative of the solvation free energy of water ΔG_S in its own liquid. The simulated ΔG_S for CSJ water is −28 kJ/mol. All other water models studied in this work give ΔG_S values from −18 to −19 kJ/mol (Table 6). The calculated ΔG_S should correspond to the ΔG_{ex} of a water tetramer (for CSJ and MARTINI W) or trimer (for SSRBK W). Direct comparison of calculated ΔG_S for a CG water representing a cluster of n water molecules to the experimental ΔG_S value (−26.5 kJ/mol calculated from the vapor pressure p_v[31] using eq 3) for water is not strictly appropriate. Baron et al.[40] found that the calculated ΔF_{ex} values for the insertion of an SPC water tetramer into SPC water and into liquid SPC tetramers are, respectively, 22 and 19 kJ/mol. These ΔF_{ex} values compare fairly with the ΔG_{ex} values of the CG water models (CSJ, MARTINI, and SSRBK) suggesting solvation of a CG water bead into its own liquid is reasonably well represented by CG water models.

When a CG water represents more than one water molecule, direct comparison of calculated ΔH_{vap} to the experiment requires a theoretical correction. Presented in Table 6 are ΔH_{vap} values directly calculated from applying eq 2a to MD simulations for the CG water models. Enthalpy of vaporization for water can then be estimated from the simulated ΔH_{vap} for CG water and the binding energy correction term of eq 2b. This has been explained in the Computational Method section. The so calculated ΔH_{vap} values for water are 43.6, 41.5, and 33.2 respectively by using the CG ΔH_{vap} values (Table 6) for CSJ W4, MARTINI W, and SSRBK W. The agreement of these calculated ΔH_{vap} values for water by using eq 2b with the experimental value, 44 kJ/mol,[30] indicates that the CG ΔH_{vap} values (Table 6) are reasonably respresented by the CG water models.

The diffusion coefficients D, calculated from the slopes of the mean square displacements (MSD) in the long time limit using the Einstein relation $\langle \Delta r(t)^2 \rangle = 6Dt$, for the various CG water models (Table 6), range from 1.6×10^{-9} to 6.6×10^{-9} m^2/s. These calculations are all newly done by us for this paper on the various models, so that the computational conditions would be the same (except the cutoff as designed for each water model.) These CG

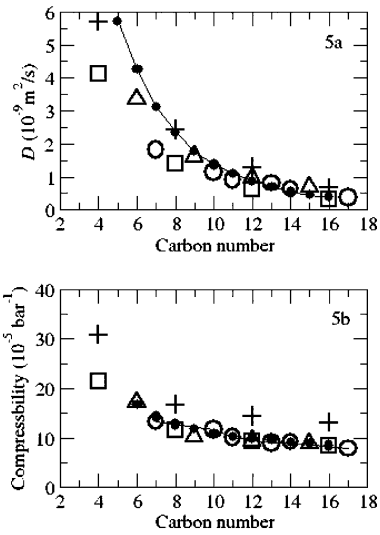


Figure 5. Self-diffusion coefficients D and isothermal compressibilities κ of alkanes at 298 K. (a) Diffusion coefficients were calculated from the slope of the mean square displacement in the long time limit. (b) Isothermal compressibilities were calculated from the relation $\kappa = (\sigma_V^2)/(kT\langle V\rangle)$, where σ_V is the volume (V) fluctuations in an NPT ensemble. Symbols for a and b: triangle, 3-site CG models; square, 4-site CG; circle, scaled CG models; filled circle, experiment.[47]

results are all on the same order of magnitude as the experimental D value (2.3×10^{-9} m^2/s)[41] for water at 298 K. Seeking precise correspondence with the experiment in this case is perhaps not meaningful, since the entity for which the experimental determination is made (single water molecules) does not exist in the CG models. The calculated diffusion coefficients, as listed in Table 6, for CSJ, MARTINI, and SSRBK CG waters were not renormalized by a factor of 4 (for CSJ and MARTINI water models) or 3 (for SSRBK), in contrast to the renormalizations done by Marrink et al.[28] as well as Groot and Rabone.[42] For any single-site CG water model which represents a group of more than one water molecule and is modeled to have the same density as liquid water, the fictitious n water molecules it represents are explicitly and implicitly bound via hydrogen bonding within the volume of the CG particle. In other words, the fictituous water molecules within the CG water model go wherever it goes. Jalabert and Das Sarma[43] have shown the diffusion coefficient of bound particles is the same as that of the center of mass of the particles.

Isothermal compressibilities κ for the water models were calculated from the equation[44] $\kappa = \sigma_V^2/kT\langle V\rangle$, where σ_V is the root-mean-square volume (V) fluctuation in an NPT ensemble. The κ values for CSJ, SSRBK, and MARTINI water (Table 6) are about 4, 3, and 2 times larger than the experimental value[30] (4.5×10^{-5} bar^{-1}) of water, respectively. We note that Marrink et al.[28] reported a smaller value of 6×10^{-5} bar^{-1} for the MARTINI water. Since this trend in the calculated κ relates to the trend in the hardness of the applied interaction potential in the short - range region, we conclude that the softer the potential in the short-range region, the larger the σ_V which in turn results in a larger κ value.

In summary, from the standpoint of coarse graining, it is appropriate to compare CG water properties to experiments for intensive properties (density, surface tension) and thermodynamic properties for equilibrium processes such as free energies of partitioning between two phases. Moreover, direct comparison of CG water results to experiment is also possible for a change-of-state thermodynamic property such as ΔH_{vap} provided a correction is made corresponding to the internal energy of the CG particle.

4.2. Alkanes. *Thermodynamic Properties.* For 3:1 and 4:1 mapped CG alkanes, experimental surface tension was used as a guide to properly choose the value of the Morse α parameter. The nonbonded CG parameters ε and R_0 were fitted against experimental ΔH_{vap} and bulk density. The scaled CG parameters were calculated as described in the Computational Methods section. The simulated

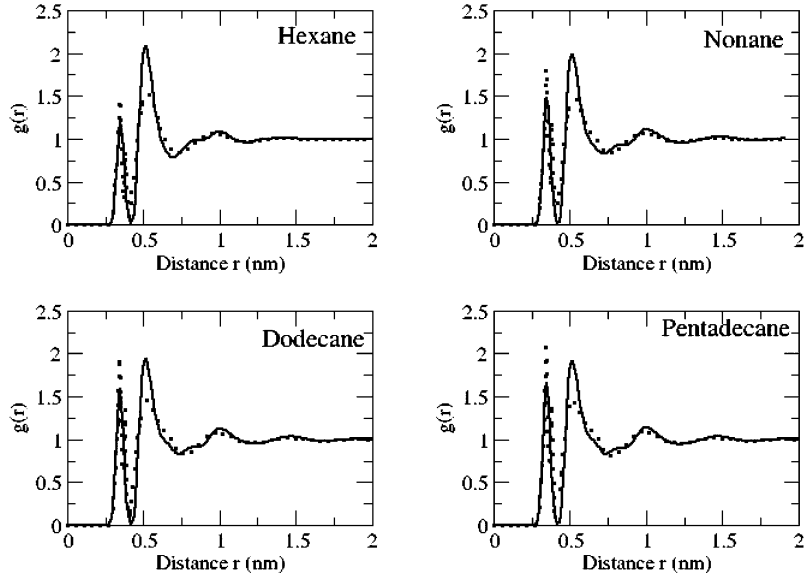


Figure 6. Comparison of radial distribution functions (RDF) for 3-site CG alkanes and corresponding mapped atomistic data. Solid line: CG RDF. Dotted line: mapped atomistic RDF.

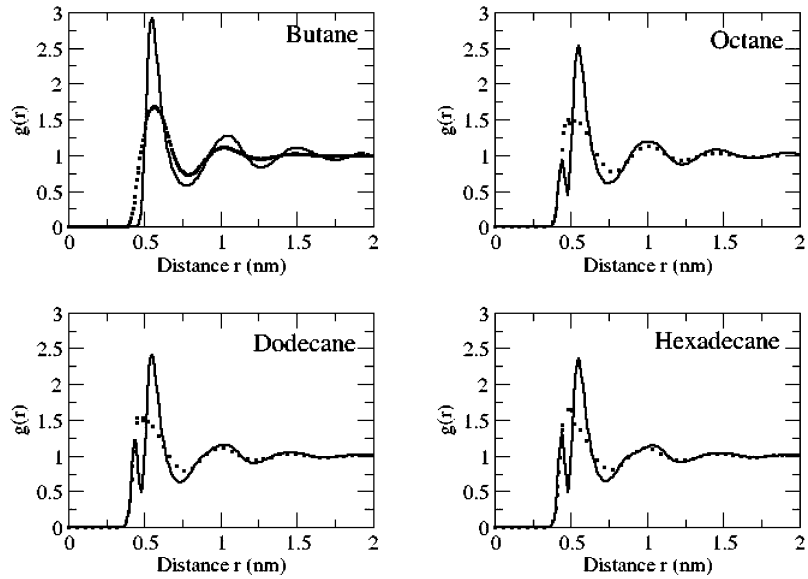


Figure 7. Comparison of radial distribution functions (RDF) for 4-site CG alkanes and corresponding mapped atomistic data. Solid line: CG RDF. Dotted line: mapped atomistic RDF.

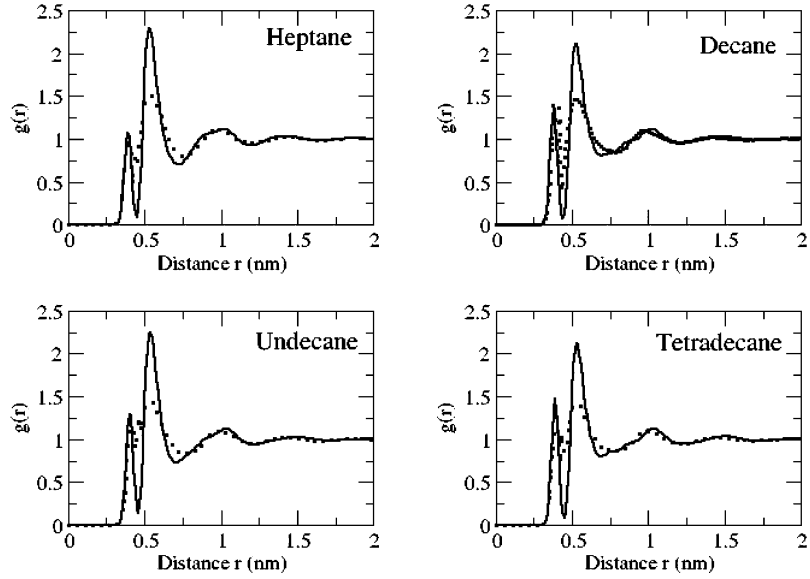


Figure 8. Comparison of radial distribution functions (RDF) for scaled CG alkanes and corresponding mapped atomistic data. Solid line: CG RDF. Dotted line: mapped atomistic RDF.

results are shown in Figure 4. Simulated ΔH_{vap} (Figure 4a) and specific volumes (Figure 4b) of CG alkanes are in good agreement with experimental data.[30] The absolute mean deviation of calculated γ from experimental values is 1.9 mN/m (Figure 4c). The ΔG_S values, as listed in Table 7, calculated by using the TI method (eq 5) agree well with those obtained from simulations of alkane vapor−liquid interfaces (eq 3). Shown in Figure 4d are the calculated ΔG_S for CG alkanes, which were not directly targeted in the parametrization process. They are systematically more negative as compared to the values based on log p measurements[31] (eq 3) for long-chain alkanes. The absolute mean deviation of CG ΔG_S from experimental values is ca. 4 kJ/mol. Although the MARTINI CG force field is less satisfactory in predicting ΔH_{vap} and specific volumes of alkanes, it delivers results of ΔG_S and γ for alkanes in good agreement with experiments.

The results of the pulsed-gradient spin−echo NMR method[45] of measuring self-diffusion establish that the diffusion coefficients D of liquid alkanes decrease nonlinearly as chain length increases,[46] as shown in Figure 5a. It can be seen from the same figure that our CG parameters also reproduce the same trend on this scale. We note that reported D values for MARTINI hexadecane by Baron et al.[40] and Winger et al.[16] are very different. Placing the actual mass on each CG site instead of the uniform value of 72 which is used in the MARTINI force field, we calculated the D values for MARTINI alkanes. The results, in good agreement with the experiment, are also shown in Figure 5a for comparison. Simulated κ values over the whole range of alkanes studied

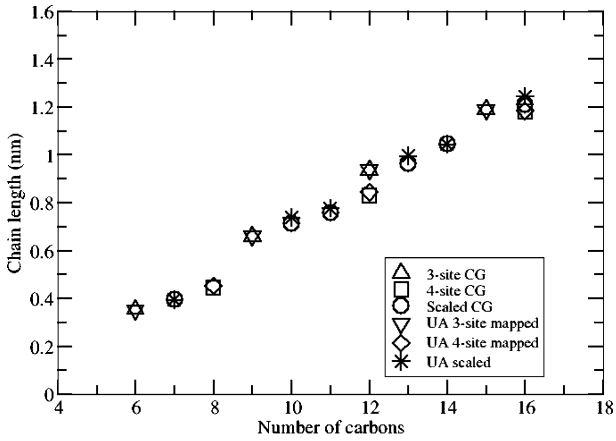


Figure 9. Comparison of mean chain lengths of CG alkanes to atomistic alkanes. Atomistic coordinates for CG mapping were taken from ref 17.

in this work agree well with the experimental results of Díaz Peña and Tardajos[47] as can be seen from Figure 5b.

In order to investigate the effects of taking torsion potentials into account on calculated energetics, CG-MD simulations for bulk CG dodecane, pentadecane (3-site models), and hexadecane (4-site model) were also performed by using the same standard parameter set (from Tables 1−3) with the torsion potential $V_d(\phi) = K_d(1 + \cos(\delta)\cos(m\phi))$ taken into account, where K_d is the dihedral force constant, δ is π, and $m = 1$.The K_d values for the 3-site and 4-site models are 0.75 and 1.2 kJ/mol, respectively. They were parametrized by matching the mean and standard deviation of torsion angles from CG-MD results with those from atomistic data mapped onto the corresponding CG models. We report in the Supporting Information that the calculated physical and thermodynamic properties for CG dodecane, pentadecane, and hexadecane are hardly influenced by the applied torsion potentials.

Structural Properties. The radial distributions (RDFs) for the CG alkanes with comparisons to corresponding mapped atomistic RDFs are shown in Figures 6−8. To compare the CG and atomistic data, the latter were analyzed using the mapping procedure as described in the Computational Method section. In these figures, the locations of the first peaks (except that for butane) correspond to the *ideal* bond lengths of the corresponding CG alkanes, which were obtained from the atomistic data, and the second peaks correspond to the inter-CG site distances. As can be seen in these figures, the structures of CG alkanes reproduce RDFs in good agreement with the atomistic data, to the extent permitted by the coarse graining.

The mean chain length (end-to-end distance) as a function of alkane carbons is plotted in Figure 9. The atomistic data, analyzed using the CG mapping procedure as described previously, are in excellent agreement with the CG data. The chain length distributions for the CG and atomistic data (Figure 10), as exemplified by the 3-site mapped *n*-alkanes (hexane, nonane, doedecane, and pentadecane), also agree well with each other. For a closer look at the agreement, we compared the CG bond angle distributions of the same series of CG alkanes (Figure 11) with those mapped from their corresponding atomistic data.

The consequence of not taking torsion potential into account is a uniform dihedral distribution for the CG model, while the corresponding atomistic model shows a bias toward lower torsion angles.[29] Including torsion potential in CG alkane with four or more consecutive interaction sites does improve dramatically the agreement of dihedral distributions between the CG and the atomistic models. The detailed results are shown in the online Supporting Information. Shown also in the Supporting Information is the comparison of distributions of bending angles with and without torsion potentials applied. It is seen there that taking torsion potential into account has no apparent effect in bending angle distributions.

In summary, inclusion of torsion potentials to the current CG alkane models shows no changed performance of the parameter set except for the excellent correspondence that

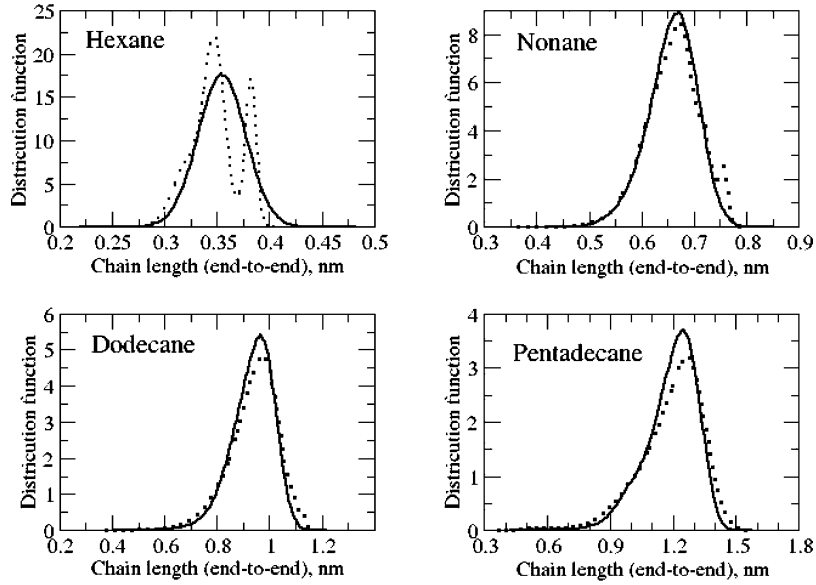


Figure 10. Comparison of chain length distributions of CG alkanes to atomistic alkanes. Solid line: CG results. Dotted line: mapped atomistic results.

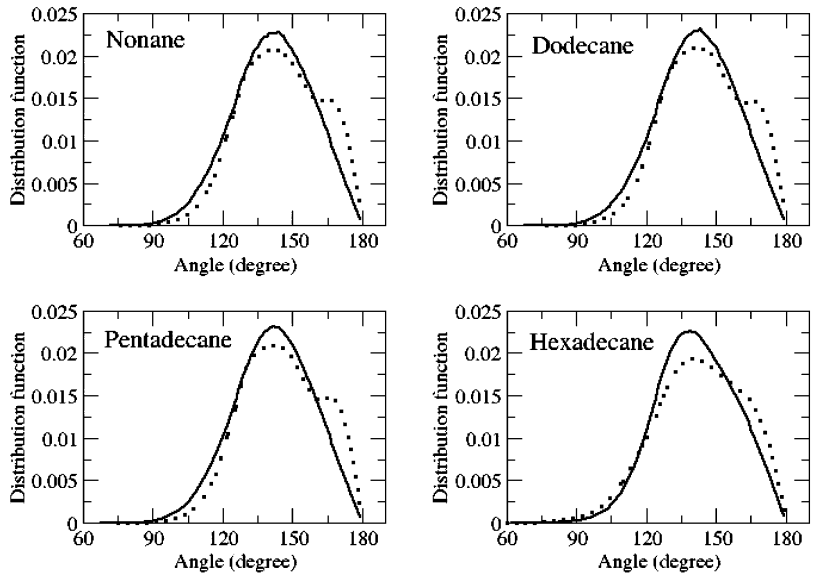


Figure 11. Comparison of bond angle distribution of CG alkanes to atomistic alkanes. Solid line: CG results. Dotted line: mapped atomistic results.

is found between CG and atomistic pseudo dihedral distributions. We note that the addition of a torsion angles makes it necessary to run simulations using a shorter Δt, so that there is a trade-off between the improved accuracy provided by this inclusion, and the capability of running longer and larger simulations without them.

4.3. Integration Time Step Analysis. Details and results for the test to evaluate how large Δt could be used in this work are presented in the Supporting Information. A criterion for a fairly accurate integration of the equations of motion is that ΔE_{tot} should be less than one-fifth of ΔE_{kin} or ΔE_{pot} in the NVE ensemble.[48−50] For both SSRBK and CSJ water models, which have softer repulsive interactions in the short-range region (Figure 1), and for CSJ hexadecane, using Δt of 40 fs is the limit to fulfill the criterion.

5. Conclusion

Coarse grained models for water and alkanes using the Morse potential (eq 1) for pairwise interactions were parametrized using the experimental ΔH_{vap}, bulk density, and surface tension to set the three adjustable parameters in the Morse form. In addition to the three experimental values mentioned above, we find that the CG models also have reasonable values for other properties. The structural properties of CG alkanes are in good agreement with those mapped from atomistic data. The application of an interaction potential softer in the short-range region allows the use of a larger integration time step for CG-MD simulations. We find that a time step of 40 fs can be safely used for the version of the force field in which torsion angle restraints are not included on the alkane chains (as they are not included in the MARTINI force field). However, if greater fidelity is desired for chain structures, a 10 fs time step is required.

Since the Morse potential can be successfully adapted for both highly polar (water) and very nonpolar (alkanes) species, we suggest it as a candidate for the general form of the nonbonded interaction in coarse-grained simulations. An additional benefit is that, due to the softer repulsive force, the Morse potential permits larger time steps than does the L-J form. It should be said the Morse potential is slightly more time-consuming (approximately 20% in our hands) per time step in our implementation (which uses table look up rather than evaluation of exponentials) than the Lennard-Jones form. However, this is more than made up by the larger time step permitted, so that on balance the Morse form is both more efficient and also more faithful in replicating a broad range of experimental data. It may also be that the Morse form calculation can be made more computationally efficient in the future.

Acknowledgment. We gratefully acknowledge support from Grant 5PN2EY016570−06 from the NIH Nanomedicine Roadmap program.

Supporting Information Available: Additional results, discussion, and methodologies inclusing tables and figures. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Müller-Plathe, F. *ChemPhysChem* **2002**, *3*, 754.

(2) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624.

(3) Spyriouni, T.; Tzoumanekas, C.; Theodorou, D.; Müller-Plathe, F.; Milano, G. *Macromolecules* **2007**, *40*, 3876.

(4) Wilson, M. R.; Stimson, L. M.; Ilnytskyi, J. M.; Hughes, Z. E. Computer Simulations Of Liquid Crystal Polymers And Dendrimers. In *Computer Simulations of Liquid Crystals and Polymers*, Pasini, P., Zannoni, C., Zumer, S., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2005; Vol. 5, pp 7−81.

(5) Prampolini, G. *J. Chem. Theory Comput* **2006**, *2*, 556.

(6) Marrink, Siewert, J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812.

(7) Monticelli, L.; Kandasamy, Senthil, K.; Periole, X.; Larson, Ronald, G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819.

(8) Loison, C.; Mareschal, M. F.; Schmid, F. *J. Chem. Phys.* **2004**, *121*, 1890.

(9) Shelly, J. C.; Shelly, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem B* **2001**, *105*, 4464.

(10) Markvoort, A. J.; Pieterse, K.; Steijaert, M. N. P.; Spijker, P.; Hilbers, P. A. J. *J. Phys. Chem. B* **2005**, *109*, 22649.

(11) Ha-Duong, T.; Basdevant, N.; Borgis, D. *Chem. Phys. Lett.* **2009**, *468*, 79.

(12) Molinero, V.; Goddard, W. A., III. *J. Phys. Chem. B* **2004**, *108*, 1414.

(13) Morse, P. M. *Phys. Rev.* **1929**, *34*, 57.

(14) Liew, C. C.; Masuhiro, M. *Chem. Phys. Lett.* **2003**, *368*, 346.

(15) Lennard-Jones, J. E. *Proc. R. Soc. London, Ser. A* **1925**, *109*, 584.

(16) Winger, M.; Trzesniak, D.; Baron, R.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1934.

(17) Chiu, S.-W.; Pandit, Sagar, A.; Scott, H. L.; Jakobsson, E. *J. Phys. Chem. B.* **2009**, *113*, 2748.

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

(19) 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. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C. GROMACS User Manual Version 4.0. http://www.gromacs.org/Documentation/Manual (accessed Jan 12, 2010).

(20) Evans, D. J.; Holian, B. I. *J. Chem. Phys.* **1985**, *83*, 4069.

(21) Parinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 182.

(22) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535.

(23) Flower, J. E.; Shaefer., H. F., III. *J. Am. Chem. Soc.* **1995**, *117*, 446.

(24) Pérez, J. F.; Hadad, C. Z.; Restrepo, A. *Int. J. Quantum Chem.* **2008**, *108*, 1653.

(25) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016.

(26) Ben-Naim, A. *J. Solution Chem.* **2001**, *30*, 475.

(27) Wescott, J. T.; Fisher, L. R.; Hanna, S. *J. Chem. Phys.* **2002**, *116*, 2361.

(28) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.

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

(30) *CRC Handbook of Chemistry and Physics*, 72nd ed.; Lide, D. R., Eds.; CRC Press. Inc.: Boca Raton, FL, 1992.

(31) *Chemical Properties Handbook: Physical, Thermodynamic, Environmental, Transport, Safety, and Health Related Properties for Organic and Inorganic Chemicals*, 1st ed.; Yaws, C. L., Ed.; McGraw-Hill: New York, 1998.

(32) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Chem. Phys.* **2003**, *119*, 7043.

(33) Buckingham, R. A. *Proc. R. Soc. London, Ser. A* **1938**, *168*, 264.

(34) Engler, E. M.; Andose, J. D.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **1973**, *95*, 8005.

(35) Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

(36) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(37) Mayo, S. L.; Olafson, B. D.; Goddard, W. A., III. *J. Phys. Chem.* **1990**, *94*, 8897.

(38) Dillen, J. M. L. *J. Comput. Chem.* **1995**, *16*, 595.

(39) Allinger, N. L.; Chen, K.; Lii, J. H. *J. Comput. Chem.* **1996**, *17*, 642.

(40) Baron, R.; Trzesniak, D.; de Vries, A. H.; Elsener, A.; Siewert, J.; Marrink, S. J.; van Gunsteren, W. F. *ChemPhysChem* **2007**, *8*, 452.

(41) Holz, M.; Heil, S. R.; Sacco, A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740.

(42) Groot, R. D.; Rabone, K. L. *Biophys. J.* **2001**, *81*, 725.

(43) Jalabert, R.; Das Sarma, S. *Phys. Rev. A.* **1988**, *37*, 2614.

(44) Herrero, C. P. *J. Phys.: Condens. Matter.* **2008**, *20*, 295230.

(45) Stejskal, E. O.; Tanner, J. E. *J. Chem. Phys.* **1965**, *42*, 288.

(46) von Meerwall, E.; Beckman, S.; Jang, J.; Mattice, W. L. *J. Chem. Phys.* **1998**, *108*, 4299.

(47) Díaz Peña, M.; Tardajos, G. *J. Chem. Thermodyn.* **1978**, *10*, 19.

(48) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Phys.* **1977**, *34*, 1311.

(49) Berendsen, H. J. C.; van Gunsteren, W. F. Practical algorithms for dynamic simulations. In *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*; Ciccotti, G., Hoover, W. G., Eds.; North-Holland: Amsterdam, 1986; Course 97, p 43.

(50) Berendsen, H. J. C.; van Gunsteren, W. F. Molecular dynamics simulation: Techniques and approaches. In *Molecular Liquids—Dynamics and Interactions*; Barnes, A. J., Orville-Thomas, W. J., Yarwood, J., Eds.; Reidell: Dordrecht, The Netherlands, 1984; Vol. 47, pp 5−500.

CT900475P

JCTC Journal of Chemical Theory and Computation

Applications of Screened Hybrid Density Functionals with Empirical Dispersion Corrections to Rare Gas Dimers and Solids

Kazim E. Yousaf and Edward N. Brothers*

Science Program, Texas A&M at Qatar, Texas A&M Engineering Building, Education City, Doha, Qatar

Received October 11, 2009

Abstract: An empirical dispersion correction is added to the range-separated hybrid density functionals HSE and HISS via parametrization versus a standard test bed of weakly bound complexes. The performance of the resulting HSE-D and HISS-D functionals is evaluated by calculating the equilibrium bond length, harmonic frequency, and dissociation energy for a number of rare gas dimers, and the lattice constants, band gaps, and sublimation energies of the rare gas solids. Both HSE-D and HISS-D are shown to provide accurate results for both molecules and extended systems, suggesting that the combination of a screened hybrid functional with an empirical dispersion correction provides an accurate, widely applicable method for use in solid-state and gas-phase electronic structure theory.

1. Introduction

Kohn−Sham density functional theory[1] (KS-DFT) continues to be a hugely popular and successful electronic structure method for a number of reasons. With modern functionals, DFT offers the tantalizing combination of accurate results at a fraction of the computational cost required for a correlated ab initio calculation. Many of today's most useful functionals belong to the class of so-called hybrid functionals, mixing a fraction of (nonlocal) Hartree−Fock exchange with conventional local exchange functionals. Hybrid functionals offer significant improvements in accuracy over those not containing exact exchange and must be used to obtain even qualitatively correct results for a number of properties of both molecules and solids. This presents a problem in the case of the latter, however, because exact exchange exhibits slow spatial decay, and, as such, traditional hybrid DFT calculations on solids are generally unfeasible. In addition, many functionals (especially older functionals) have difficulty with molecular and periodic systems in which dispersion interactions play an important role. Such interactions are known to arise from long-range electron correlations, which are inadequately modeled by the majority of functionals.

Attempts have been made, in recent years, to overcome both of these issues. In range-separated hybrid functionals, the Coulomb operator is partitioned into two or more distance ranges, and with appropriate parametrization the fraction of exact exchange included is allowed to vary as a function of interelectronic separation r_{12}. The functional of Heyd, Scuseria, and Ernzernhof[2,3] (HSE, also known as HSE06) is an early example of such a functional and was designed to include exact exchange only at small r_{12}, thus ameliorating the computational expense of modeling solids. HSE is among the most accurate density functionals for band gap calculations in solids due to this short-range Hartree−Fock exchange and has also been shown to yield good structural and energetic data for extended systems.[4−7] However, its performance in molecular calculations is not as impressive because the exchange potential now lacks the correct asymptotic behavior. The LC-ωPBE functional[8,9] is designed in a similar way to HSE but, conversely, includes only long-range exact exchange. This functional consequently outperforms HSE for many molecular properties, but is unsuitable for calculations on extended systems. Henderson, Izmaylov, Scuseria, and Savin (HISS) developed a functional[10] in an attempt to combine the best features of the HSE functional in calculations on solids with those of the LC-ωPBE functional in molecular calculations. In the HISS functional,

* Corresponding author e-mail: ed.brothers@qatar.tamu.edu.

10.1021/ct900536n
Published on Web 02/11/2010

the Coulomb operator is split not only into short-range and long-range components, as in HSE and LC-ωPBE, but also into a middle-range part. HISS is parametrized so that exact exchange is only computed in this middle r_{12} range and has been shown to give only slightly less accurate results on solids than HSE, while yielding far superior accuracy for finite systems.[11]

Regardless of the importance of exact exchange, however, it cannot improve the description of a system in which dispersion interactions are important. Many different solutions to this problem have been proposed, but in this work we focus on the popular empirical dispersion correction of Grimme, commonly known as the DFT-D method.[12,13] In this conceptually simple approach, a pairwise empirical correction, which depends on a single, optimized functional-dependent scaling parameter, is added to the KS-DFT energy. There are obvious limitations to this approach, as a semiempirical correction can only improve an already a reasonable semilocal interaction curve, and not all dispersion interactions are pairwise.[14] However, this correction is extremely quick to calculate as compared to the self-consistent field step of a KS-DFT calculation, even for large systems, and DFT-D has yielded excellent geometries and interaction energies for dispersion bound complexes.[12,13] Recently, the DFT-D method has been extended to the domain of solid-state electronic structure calculations by a number of groups with encouraging results.[15−17]

By combining a screened hybrid functional with an empirical dispersion correction, it is hoped that an accurate universal functional that could be applied to both molecules and solids could be constructed. In the present study, optimal s_6 values are determined for the HSE and HISS functionals against a standard set of weakly bound complexes, and the performance of the resultant HSE-D and HISS-D functionals is assessed by calculating the geometry, harmonic frequency, and dissociation energy for a number of rare gas dimers. With our implementation of DFT-D under periodic boundary conditions, these functionals are then used to calculate the lattice parameters, sublimation energies, and band gaps of the dispersion bound rare gas solids. Note that these systems are vastly different from those in our training set and were selected to demonstrate the transferability of our parametrization and general applicability of our corrected functionals.

2. Theory

2.1. Range-Separated Hybrid Density Functionals. A recent overview of the theory and applications of range-separated hybrids was given by Henderson et al.;[18] the interested reader is referred to this paper, but a brief summary is given here for the sake of convenience. Following the ideas of Savin,[19,20] in this type of functional the Coulomb operator r_{12}^{-1} is typically partitioned into a short-range and a long-range component:

$$\frac{1}{r_{12}} = \frac{\mathrm{erfc}(\omega r_{12})}{r_{12}} + \frac{\mathrm{erf}(\omega r_{12})}{r_{12}} \tag{1}$$

where the first and second terms on the right-hand side contain the short- and long-range (SR and LR) parts of the exchange, respectively, the complementary error function is defined as $\mathrm{erfc}(\omega r) = 1 - \mathrm{erf}(\omega r)$, and ω is an adjustable parameter that controls the definition of the two ranges. In the Heyd−Scuseria−Ernzerhof (HSE) screened hybrid functional, the first of these range-separated hybrids considered in this work, $\omega = 0.11\ a_0^{-1}$. The error function is not the only suitable function for use in this partitioning, but has the desirable property that the requisite integrals are easy to calculate analytically. HSE is based on the PBE0 hybrid functional,[21,22] and the HSE exchange-correlation energy is given by

$$E_{\mathrm{xc}}^{\mathrm{HSE}} = E_{\mathrm{xc}}^{\mathrm{PBE}} + \frac{1}{4}(E_{\mathrm{x}}^{\mathrm{HF,SR}} - E_{\mathrm{x}}^{\mathrm{PBE,SR}}) \tag{2}$$

with the mixing coeffecient 1/4 obtained from perturbation theory.[23] HSE is thus equivalent to PBE0 for $\omega = 0$ and approaches PBE as $\omega \rightarrow \infty$. Although it has been shown to yield accurate band gaps, lattice constants, and bulk moduli in solids,[4−7] for example, it performs less well for molecular thermochemistry and reaction barriers due to the incorrect behavior of the exchange potential, HSE only includes a significant fraction of exact exchange for $r_{12} \leq 1/\omega \approx 4.8$ Å. On the basis of the long-range correction scheme of Iikura and co-workers,[24] the LC-ωPBE functional was introduced by Vydrov and Scuseria[8,9] to address this shortcoming and, opposite to HSE, exclusively contains Hartree−Fock exchange at large r_{12}. The LC-ωPBE energy is given by

$$E_{\mathrm{xc}}^{\mathrm{LC}-\omega\mathrm{PBE}} = E_{\mathrm{x}}^{\mathrm{PBE,SR}} + E_{\mathrm{x}}^{\mathrm{HF,LR}} + E_{\mathrm{c}}^{\mathrm{PBE}} \tag{3}$$

with the parameter $\omega = 0.4\ a_0^{-1}$ ($r_{12} \approx 1.3$ Å). Note that the Coulomb-attenuated method (CAM) functionals of Yanai and co-workers[25] had previously been introduced to achieve a similar goal. CAM-B3LYP, for example, yields hugely improved charge transfer excitations as compared to B3LYP but differs from the LC hybrid functionals in that it still contains a small fraction of exact exchange at small r_{12}. While LC-ωPBE improves upon HSE for a number of molecular properties, the inclusion of long-range exact exchange makes the functional unsuitable for calculations on solids. The motivation for the development of the HISS (Henderson−Izmaylov−Scuseria−Savin) functional[10] was to introduce a functional that benefits from the increased accuracy afforded by including exact exchange at long-range, but that is still suitable for studies of extended systems. In the HISS functional, the Coulomb operator is split into three length ranges:

$$\frac{1}{r_{12}} = \underbrace{\frac{\mathrm{erfc}(\omega_{\mathrm{SR}} r_{12})}{r_{12}}}_{\mathrm{SR}} + \underbrace{\frac{\mathrm{erfc}(\omega_{\mathrm{LR}} r_{12}) - \mathrm{erfc}(\omega_{\mathrm{SR}} r_{12})}{r_{12}}}_{\mathrm{MR}} + \underbrace{\frac{\mathrm{erf}(\omega_{\mathrm{LR}} r_{12})}{r_{12}}}_{\mathrm{LR}} \tag{4}$$

where MR denotes the additional middle-range length scale. As for HSE, exact exchange contributions for large r_{12} must be neglected to ensure the applicability of the functional to solids, but as in LC-ωPBE, short-range exact exchange is also omitted. That is, HISS contains exact exchange only in the middle range. Thus, we can write the total HISS exchange-correlation energy as

$$E_{xc}^{HISS} = E_{xc}^{PBE} + \frac{3}{5}(E_x^{HF,MR} - E_x^{PBE,MR}) \quad (5)$$

where the mixing coefficient 3/5 has been shown to be thermochemically optimal. The definition of a third range requires two values of the parameter ω, which have been determined as $\omega_{SR} = 0.84\ a_0^{-1}$ ($r_{12} \approx 0.6$ Å) and $\omega_{LR} = 0.20\ a_0^{-1}$ ($\sim$2.6 Å). Note that both HSE and LC-ωPBE are special cases of HISS-type functionals and can be defined within the framework of a three-range Coulomb operator with appropriate parameters.

2.2. Empirical Dispersion Corrections. Following the prescription of Grimme,[12] the DFT-D energy can be written as

$$E_{DFT-D} = E_{DFT} + E_{disp} \quad (6)$$

For an N-atom system, E_{disp} is defined as the simple pairwise sum:

$$E_{disp} = -s_6 \frac{1}{2} \sum_{i \neq j} \frac{C_6^{ij}}{R_{ij}^6} f(R_{ij}) \quad (7)$$

where s_6 is the global scaling parameter, which depends only on the functional in use, C_6^{ij} is the geometric mean of the individual atomic dispersion coefficients, R_{ij} is interatomic distance for a given pair, and f is the damping function:

$$f = \frac{1}{1 + \exp\left[-d\left(\frac{R_{ij}}{R_r} - 1\right)\right]} \quad (8)$$

where R_r is the sum of van der Waals radii for a given pair. Accurate dispersion coefficients and careful damping ($d = 20$ has been shown to be optimal)[13] are vital to ensure the correction applies only to medium-range R_{ij}, where the dispersion forces are greatest. The derivatives $\partial E_{disp}/\partial R_{ij}$ and $\partial^2 E_{disp}/\partial R_{ij}^2$ are straightforward to determine and implement, enabling DFT-D geometry optimizations and frequency calculations to be performed for gas-phase molecules.

The DFT-D energy expression can be modified for extended systems by adding a sum over lattice vectors; then a correction to E_{DFT-D}, taking periodic boundary conditions (PBC) into account, can be computed as

$$E_{DFT-D}^{PBC} = E_{DFT} + E_{disp}^{g=0} - \frac{s_6}{2} \sum_{i=1}^{N} \sum_{g \neq 0} \sum_{j=1}^{N} \frac{C_6^{ij}}{R_{ij,g}^6} f(R_{ij,g}) \quad (9)$$

The term $E_{disp}^{g=0}$ is the dispersion interaction between atoms in the unit cell, while the final term is calculated between atoms i in the unit cell with atoms j in the image cells, where the image cells are generated by integer multiples g of each lattice vector and the factor of 1/2 is to avoid double counting. The gradient of the PBC dispersion energy has also been derived and implemented and is rendered only slightly more complicated than for the molecular case because of the need to determine the derivatives with respect to lattice vectors as well as atoms.[26]

There are previous reports of DFT-D for periodic systems in the literature. In one of the earliest such implementations, Ortmann et al.[15] performed periodic DFT-D calculations using plane-wave basis sets, with the PW91-D functional yielding mixed results for a wide variety of weakly bound systems. More recently, Kerber et al.[16] obtained accurate results for the interlayer spacing and binding energy in graphite and vanadia, as well as good reaction energies for the adsorption of organic molecules on silica and zeolites using the PBE-D functional. In addition, the B3LYP- D* functional of Civalleri et al.[17] has been shown to give good structural and energetic data for a representative set of molecular crystals. This empirically corrected functional is not quite the same as the B3LYP-D functional as defined within the framework of Grimme, as the atomic van der Waals radii are scaled instead of the entire dispersion correction. That is, s_6 in eqs 7 and 9 is equal to unity, but the van der Waals radii R_r in eq 8 are multiplied by a parameter s_R (equal to 1.3 for hydrogen and 1.05 for heavy atoms). This is similar to the approach of Jurečka et al.[27] except that they use a single value of s_R, optimized for each functional and basis set combination, for the whole periodic table.

Unlike the implementation of Civalleri et al., we do not have a fixed distance-based cutoff, but rather add contributions to E_{disp} from ever larger shells of matter until some convergence criterion κ is reached. This approach was chosen due to the unexpectedly long-range of non-negligible dispersion contributions found during testing, and the strong system dependence of such a cutoff. Typically, $\kappa = 10^{-10}\ E_h$ was used for both the energy and the gradient. This is tighter than is strictly necessary as this far exceeds the accuracy of the DFT quadrature, for example, but guarantees the smoothness of our DFT-D potential energy surfaces.

3. Computational Details

All calculations were performed using a development version of the Gaussian electronic structure program.[28] A number of basis sets were considered for use in this work. After our initial tests, we chose to use the def2-TZVPP basis sets[29] for the PBC calculations, whereas for the dimer calculations the aug-cc-pVQZ basis set was selected for Ne[30] and Kr,[31] while the aug-cc-pV(Q+d)Z[32] and aug-cc-pVQZ-PP[33] basis sets were used for Ar and Xe, respectively. These basis sets were used in conjunction with the small-core pseudopotential of Peterson et al.[33] for xenon in the solid and molecular calculations. For the parametrizations, the 6-311++G(3df,3pd) basis set[34] was used. Because the DFT-D method is best used with large, flexible basis sets,[12] this basis set was chosen to ensure the transferability of our parametrization while enabling the requisite SCF calculations to be performed in reasonable time. The counterpoise correction of Boys and Bernardi[35] was not computed for any of the systems studied, as basis set superposition errors for these systems have previously been shown to be very small with triple- or quadruple-ζ basis sets.[36] Tight and very tight convergence criteria were used throughout, for the SCF procedure and geometry optimizations, respectively. Furthermore, it was found during the initial stages of this work that very dense integration grids must be employed to achieve sufficiently accurate results for both the solid and the molecular

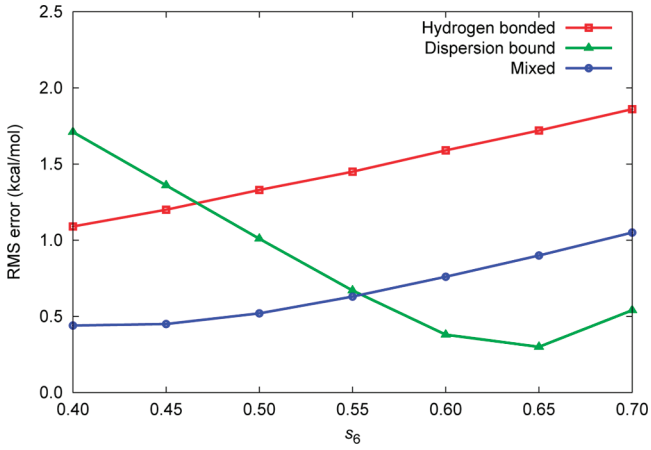


Figure 1. Plot of the rms errors in the HSE/6-311++G(3df,3pd) interaction energies for each type of complex in the S22 set against s_6, the global DFT-D scaling parameter.

calculations (especially the latter), and a grid with 199 radial and 974 angular points per atom was selected (for comparison, the "ultrafine" grid in Gaussian is a pruned grid with 99 radial and 590 angular points). This point is discussed further in section 5, as the choice of integration grid was found to have a significant effect, especially on some of the harmonic frequency calculations.

Although the main focus of this work was to evaluate the performance of the screened hybrid functionals HSE and HISS combined with an empirical dispersion correction, a number of other functionals were selected for comparison. In addition to BLYP,[37−39] PBE,[40] and TPSS[41] (both with and without the empirical correction), the M06-L functional[42] of Zhao et al. and Grimme's B97-D functional[13] were also selected. Of all of the recent Minnesota functionals, M06-L was chosen as it does not contain any nonlocal exchange, making it suitable for calculations on solids as well as molecules.[43]

4. Parametrization

As noted above, eq 7 contains one adjustable parameter, s_6, which depends only on the functional being used. Optimal values for this parameter have been determined for a number of popular functionals, but not for the screened functionals under consideration here. We determined the optimal s_6 values for the HSE and HISS functionals by minimizing the root-mean-square (rms) error of the interaction energies in the S22 set of molecular complexes. The S22 test set was proposed by Jurečka et al.[44] and contains three types of complex: hydrogen bonded (7 complexes), dispersion bound (8), and mixed (7). During the validation of our parametrization on functionals for which the optimized s_6 value has been determined, it was observed that DFT-D can degrade the performance of a given functional for hydrogen-bonded systems. In other words, the most accurate DFT-D interaction energies for these complexes are obtained when $s_6 = 0$ for a number of functionals; we note that this issue influenced the parametrization procedure of Jurečka and co-workers.[27] The need for a balanced test set is illustrated in Figure 1 by the rms interaction energy errors for each type of complex in the S22 set, calculated with the HSE functional and graphed against s_6, which clearly shows that simply minimizing the error for the dispersion bound subset would result in significantly larger errors for the other types of complexes.

Table 1. The rms Errors in the Interaction Energy (kcal/mol) in the S22 Set Calculated Using the HSE and HISS Functionals and the 6-311++G(3df,3pd) Basis Set, Both With and Without an Empirical Dispersion Correction

	HSE		HISS	
	DFT	DFT-D	DFT	DFT-D
hydrogen bonded	0.78	1.45	1.63	0.62
dispersion bound	4.56	0.67	4.38	0.48
mixed	1.46	0.63	1.52	0.65
overall	2.90	0.98	2.92	0.58

For both HSE and HISS, the optimal value of s_6 was determined to be 0.55. Table 1 shows the rms errors of the interaction energies for the S22 set complexes calculated with both functionals, with and without the empirical dispersion correction, and demonstrates the marked improvement of DFT-D for dispersion bound and mixed complexes. It is interesting to note that, while the increase in accuracy observed with HSE-D is a trade-off (because the errors for hydrogen-bonded systems increase), the addition of an empirical correction to HISS results in a uniform improvement for each type of complex. This suggests, prior to any detailed testing, that HISS-D in particular should provide good results across a range of weakly bound systems.

5. Applications

To assess the performance of DFT-D methods and especially the new HSE-D and HISS-D functionals for dispersion bound systems, we calculated the bond lengths (Table 2), harmonic frequencies (Tables 3 and 4), and dissociation energies (Table 5) of the rare gas dimers Ne_2, Ar_2, Kr_2, Xe_2, NeAr, NeKr, NeXe, ArKr, ArXe, and KrXe using the functionals stated above. The same methods were then used to calculate the lattice constants (Table 6), sublimation energies (Table 7), and band gaps Table 8) for the face-centered cubic structures of solid Ne, Ar, Kr, and Xe. These systems were chosen because, in addition to being bound solely by dispersion forces, they are quite different from those included in the parametrization; the S22 set features complexes of molecules containing only the atoms H, C, N, and O. Helium was excluded from our study as it does not exist as a solid at standard pressures, and our principal aim in this work is to demonstrate the applicability of HSE-D and HISS-D for both solid and molecular calculations.

For the dimers, we compare our calculated results to the experimental values of both Ogilvie and Wang[45,46] (OW) and Tang and Toennies[47] (TT) (note that these potentials are experimental in the sense that they are fits to experimental and not theoretical data). Although there is little difference between the two for equilibrium properties, the more recent TT potentials have been championed by Gerber and Ángyán,[48] and Ruzsinszky et al.[49] have recently shown that the OW potentials are actually divergent at large interatomic separation. Nonetheless, the OW potentials have been widely used as reference values in a number of theoretical studies, including the recent DFT investigations of Tao and Perdew[36]

Table 2. Equilibrium Bond Lengths of the Rare Gas Dimers (Å) Calculated Using the aug-cc-pVQZ Basis Sets

	Ne_2	Ar_2	Kr_2	Xe_2	NeAr	NeKr	NeXe	ArKr	ArXe	KrXe
HISS	3.4192	4.3244	4.6180	4.9450	3.8829	4.0590	4.3008	4.4798	4.6817	4.7957
HSE	3.1020	4.0254	4.3677	4.7805	3.5737	3.7604	3.9919	4.1980	4.4184	4.5836
M06-L	3.1496	3.9659	4.3836	4.9129	3.5618	3.8197	4.0143	4.2457	4.5086	4.7283
PBE	3.0880	3.9958	4.3459	4.7526	3.5420	3.7222	3.9409	4.1729	4.3885	4.5545
TPSS	3.3188	4.2698	4.6568	5.1624	3.7857	3.9770	4.2230	4.4713	4.7181	4.9057
B97D	3.2786	4.0428	4.1488	4.3772	3.6768	3.7832	3.9524	4.1003	4.2377	4.2729
BLYP-D	2.9327	3.8766	4.0664	4.3765	3.4222	3.5473	3.7413	3.9745	4.1472	4.2272
HISS-D	3.1196	3.9406	4.1216	4.3896	3.5495	3.6720	3.8600	4.0384	4.1984	4.2655
HSE-D	2.9312	3.8053	4.0545	4.3593	3.3818	3.5339	3.7353	3.9377	4.1179	4.2168
PBE-D	2.9004	3.7444	3.9891	4.3015	3.3300	3.4713	3.6585	3.8721	4.0459	4.1516
TPSS-D	3.0016	3.8607	4.0685	4.3429	3.4441	3.5759	3.7637	3.9735	4.1401	4.2152
OW[a]	3.0910	3.7565	4.0080	4.3627	3.4889	3.6210	3.8610	3.8810	4.0668	4.1740
TT[b]	3.0904	3.7572	4.0112	4.3657	3.4767	3.6460	3.8895	3.8895	4.0905	4.1964

[a] References 45 and 46. [b] Reference 47.

Table 3. Harmonic Frequencies of the Rare Gas Dimers (cm^{-1}) Calculated Using the aug-cc-pVQZ Basis Sets

	Ne_2	Ar_2	Kr_2	Xe_2	NeAr	NeKr	NeXe	ArKr	ArXe	KrXe
HISS	15.2	12.3	7.3	6.8	12.4	11.4	9.6	9.6	8.6	7.2
HSE	31.3	21.1	13.3	10.7	25.4	22.6	20.2	18.0	15.6	11.9
M06-L	77.8	47.8	33.3	23.8	46.6	53.2	47.9	35.2	28.8	28.0
PBE	33.9	22.8	15.1	11.6	28.5	25.3	23.9	19.0	17.6	13.5
TPSS	23.7	15.0	11.9	7.5	21.3	18.3	17.0	14.0	12.5	9.0
B97-D	31.0	22.6	21.7	23.7	29.1	25.8	26.5	23.7	24.1	23.2
BLYP-D	42.0	23.7	24.2	23.5	29.9	30.8	30.2	24.7	24.9	23.8
HISS-D	30.9	23.1	20.9	20.1	25.8	25.3	24.6	22.1	22.3	20.4
HSE-D	47.6	30.5	24.0	21.3	37.3	33.9	33.1	28.0	26.8	22.5
PBE-D	54.3	36.6	29.0	26.2	44.5	41.4	39.4	33.3	32.1	27.6
TPSS-D	43.7	30.4	24.8	23.6	36.3	34.1	32.8	27.8	26.7	24.1
OW[a]	28.5	30.9	23.6	20.9	28.2	26.2	24.3	27.9	27.1	22.7
TT[b]	29.4	32.0	24.3	21.2	28.5	25.3	22.8	28.7	26.9	22.7

[a] References 45 and 46. [b] Reference 47.

Table 4. Effect of the Integration Grid on the Harmonic Stretching Frequency of the Neon Dimer, Calculated at the TPSS/aug-cc-pVQZ Level With and Without the Boys−Bernardi Counterpoise Correction[a]

		ω_e (cm^{-1})	
grid	points per atom	without CP	with CP
pruned (75,302) "fine"	8338	46.5	46.7
pruned (99,590) "ultrafine"	23 416	41.2	34.3
(199,974)	167 528	23.7	23.5
(96,32,64)	169 984	44.8	35.8
(200,100,200)	3 440 000	25.2	23.6
(400,200,400)	27 520 000	25.6	24.5

[a] For notation, see the text.

and Zhao and Truhlar.[50] Errors are thus reported relative to the OW values, but generally the OW and TT potentials are in very close agreement for the properties studied here. We are also compelled to cite the highly accurate potentials of Aziz,[51,52] which were the preferred reference values of Ruzsinszky et al.[49] Finally, we note that the importance of the higher-order coefficients C_8 and C_{10} in the TT potentials has influenced the recent damped dispersion correction for GGA functionals of Steinmann et al.,[53] which appears to yield very promising results for a range of systems. However, these higher-order terms do not appear in the DFT-D framework of Grimme used in this work, although they have recently been derived from the exchange-hole dipole moment for the systems studied here by Becke and Johnson.[54]

Deviations from experimental values are reported as both mean signed errors (MSE) and mean unsigned errors (MUE) in Table 9 for the dimers and Table 11 for the solids. For the harmonic frequencies and dissociation energies of the dimers, these quantities are also expressed as percentages (see Table 10) because the experimental values are very small by chemical standards (dissociation energies are as low as 0.084 kcal/mol for Ne_2, rising to only 0.561 kcal/mol for Xe_2).

The uncorrected functional BLYP failed to bind any of the solids or dimers under investigation, and so results for this functional without an empirical dispersion correction are not tabulated in this section.

5.1. Dimers. *5.1.1. Bond Lengths.* Bond lengths for the rare gas dimers are given in Table 2. All of the DFT-D methods perform extremely well, and it is clear that the empirical correction is a necessity to obtain accurate geometries for these systems. Given the failure of its parent functional to bind any of the dimers at all, the accuracy of BLYP-D is especially pleasing. Without the dispersion correction PBE is the best functional, while HISS and TPSS give particularly large errors, but every uncorrected functional significantly overestimates the bond lengths in these dimers as shown by their positive mean signed errors in Table 9.

5.1.2. Harmonic Frequencies. The harmonic frequencies are listed in Table 3. B97-D is clearly the best performing functional here, with HISS-D, HSE-D, BLYP-D, and TPSS-D following close together. The empirical correction improves the results for HISS much more than for HSE, and this

Table 5. Dissociation Energies of the Rare Gas Dimers (kcal/mol) Calculated Using the aug-cc-pVQZ Basis Sets

	Ne_2	Ar_2	Kr_2	Xe_2	NeAr	NeKr	NeXe	ArKr	ArXe	KrXe
HISS	0.025	0.032	0.044	0.057	0.028	0.030	0.031	0.037	0.040	0.049
HSE	0.084	0.108	0.124	0.143	0.094	0.096	0.099	0.115	0.120	0.132
M06-L	0.162	0.142	0.103	0.136	0.171	0.167	0.193	0.129	0.144	0.122
PBE	0.122	0.145	0.163	0.181	0.135	0.140	0.145	0.153	0.158	0.171
TPSS	0.070	0.075	0.086	0.087	0.075	0.079	0.080	0.080	0.080	0.087
B97-D	0.146	0.255	0.493	0.825	0.186	0.239	0.276	0.349	0.427	0.626
BLYP-D	0.073	0.087	0.282	0.508	0.071	0.119	0.146	0.165	0.218	0.374
HISS-D	0.096	0.159	0.288	0.472	0.116	0.145	0.163	0.211	0.255	0.362
HSE-D	0.195	0.278	0.412	0.593	0.221	0.252	0.273	0.333	0.380	0.485
PBE-D	0.279	0.393	0.586	0.835	0.321	0.371	0.409	0.473	0.542	0.689
TPSS-D	0.226	0.329	0.532	0.793	0.263	0.314	0.348	0.412	0.477	0.638
OW[a]	0.084	0.285	0.400	0.561	0.134	0.142	0.147	0.361	0.375	0.464
TT[b]	0.084	0.285	0.400	0.562	0.132	0.141	0.141	0.333	0.373	0.464

[a] References 45 and 46. [b] Reference 47.

Table 6. Lattice Constants of the Rare Gas Solids (Å) Calculated with the def2-TZVPP Basis Set

	Ne	Ar	Kr	Xe
HISS	4.4574	5.8372	6.3487	6.9689
HSE	4.3768	5.7486	6.2896	6.9561
M06-L	4.2347	5.2605	5.5465	6.0045
PBE	4.3507	5.7436	6.3108	6.9978
TPSS	4.5523	6.1456	6.8171	7.6287
B97-D	4.4574	5.6027	5.7183	6.0507
BLYP-D	4.0464	5.2665	5.5808	6.0296
HISS-D	4.2214	5.3789	5.6688	6.0481
HSE-D	4.1604	5.3732	5.7123	6.1142
PBE-D	4.1203	5.2967	5.6271	6.0468
TPSS-D	4.2315	5.4354	5.6604	5.9807
expt.	4.464[a]	5.311[b]	5.67[c]	6.132[d]

[a] Reference 58. [b] Reference 59. [c] Reference 60. [d] Reference 61.

Table 7. Sublimation Energies of the Rare Gas Solids (J/mol) Calculated with the def2-TZVPP Basis Set

	Ne	Ar	Kr	Xe
HISS	2044	1332	1334	1319
HSE	2959	2459	2472	2497
M06-L	8471	11 119	12 640	20 192
PBE	4281	3120	2943	2842
TPSS	2834	1721	1509	1343
B97-D	6140	7928	14 687	25 462
BLYP-D	7533	7043	13 188	20 753
HISS-D	5054	6178	10 030	15 376
HSE-D	6270	7466	10 947	15 493
PBE-D	8988	10 298	15 112	21 536
TPSS-D	7918	8991	14 729	23 012
expt.	1933[a]	7732[b]	11 158[b]	15 839[b]

[a] Reference 62. [b] Reference 63.

Table 8. Band Gaps of the Rare Gas Solids (eV) Calculated with the def2-TZVPP Basis Set

	Ne	Ar	Kr	Xe
HISS	26.89	20.30	16.58	13.35
HSE	22.82	16.98	13.76	10.77
M06-L	21.29	16.75	12.60	8.43
PBE	19.43	15.06	12.22	9.57
TPSS	21.72	16.27	12.63	10.35
B97-D	21.01	15.39	11.90	8.16
BLYP-D	16.79	14.78	11.20	7.96
HISS-D	24.85	20.15	15.99	11.71
HSE-D	20.98	16.94	13.22	9.32
PBE-D	17.52	14.97	11.46	8.01
TPSS-D	19.10	15.70	12.14	8.46
expt.[a]	21.56	15.76	14.00	12.13

[a] Reference 64.

Table 9. Errors in the Bond Lengths, Harmonic Frequencies, and Dissociation Energies of the 10 Rare Gas Dimers Studied[a]

	r_e (Å)		ω_e (cm^{-1})		D_e (kcal/mol)	
	MSE	MUE	MSE	MUE	MSE	MUE
HISS	0.5196	0.5196	−16.0	16.0	−0.258	0.258
HSE	0.2491	0.2491	−7.0	7.6	−0.184	0.184
M06-L	0.2980	0.2980	16.2	16.2	−0.148	0.185
PBE	0.2192	0.2198	−4.9	6.1	−0.144	0.152
TPSS	0.5178	0.5178	−11.0	11.0	−0.215	0.215
B97-D	0.1560	0.1560	−0.9	2.7	0.087	0.095
BLYP-D	0.0001	0.0838	1.7	4.3	−0.091	0.091
HISS-D	0.0844	0.0846	−2.5	3.0	−0.069	0.075
HSE-D	−0.0237	0.0729	4.5	4.6	0.047	0.054
PBE-D	−0.0846	0.0846	10.4	10.4	0.195	0.195
TPSS-D	0.0075	0.0668	4.4	4.6	0.138	0.138

[a] Errors are given relative to the experimental results of Ogilvie and Wang.

appears to be due to the poor quality of the HSE-D frequencies for neon-containing dimers. Indeed, if only the Ar, Kr, and Xe dimers are considered, the accuracy of the HSE-D frequencies is spectacular (with a MSE and MUE equal to 0.0 and 0.3 cm^{-1}, respectively). Conversely, for the uncorrected HISS and HSE functionals, the errors in the calculated homonuclear dimer frequencies are far larger for the heavier elements. Of course, the effect of the empirical correction is most pronounced for the BLYP functional, with BLYP-D once again giving very good results here. The errors observed with the M06-L functional are surprisingly large, with the neon dimer proving especially troublesome.

We note that our frequencies differ significantly from those reported by Tao and Perdew,[36] who investigated the He, Ne, Ar, and Kr dimers using the LSDA, PBE, and TPSS functionals. Although we used the same basis set (except for Ar), there is a substantial discrepancy between some of our numbers, and this can be attributed solely to the choice of integration grid. We note that the need to use dense integration grids for GGA-type calculations on dispersion-bound systems has recently been highlighted and explained in great detail by Johnson et al.[55] The effects of using a larger grid, (199,974) in this work, are not uniform, but are especially pronounced in a few cases. Our value of ω_e for

Table 10. Percentage Errors in the Harmonic Frequencies and Dissociation Energies of the 10 Rare Gas Dimers Studied[a]

	ω_e		D_e	
	MSE	MUE	MSE	MUE
HISS	−62	62	−84	84
HSE	−28	30	−51	51
M06-L	60	60	−23	57
PBE	−20	24	−31	40
TPSS	−43	43	−63	63
B97-D	−2	10	37	40
BLYP-D	7	16	−30	30
HISS-D	−9	11	−17	22
HSE-D	17	17	36	38
PBE-D	39	39	97	97
TPSS-D	17	17	69	69

[a] For the harmonic frequencies, errors are given relative to the experimental results of Ogilvie and Wang as in Table 9.

Ne_2, calculated with TPSS, differs by 18.2 cm^{-1} from Tao and Perdew's value (over 60% of the experimental value). Focusing on this case, in particular, we chose to investigate this effect further by calculating ω_e with a variety of grids both with and without the counterpoise correction.[35] The results are tabulated in Table 4 and show that the grids usually used in benchmark calculations are inadequate in this case. The harmonic frequency calculated with the (199,974) grid used throughout this work agrees well with the value obtained with the (400,200,400) spherical product grid, with the latter containing 27.5 million points per atom. Note also that the magnitude of the basis set superposition error decreases with ever denser quadrature grids.

5.1.3. Dissociation Energies. Table 5 shows the dissociation energies for the 10 dimers considered here. All functionals yield significant errors in at least one case, but the best results are clearly obtained with the HISS-D functional. As for the harmonic frequencies, the effect of the dispersion correction here is varied and results in a far greater improvement for HISS than HSE, for example, and once more this seems to be due to the poor accuracy for the dimers containing neon. If these four dimers are excluded from the statistics, the HSE-D functional again yields astonishingly high accuracy (MSE of 0.006 kcal/mol, MUE equal to 0.017 kcal/mol). The effect of the correction on TPSS is limited, while it substantially degrades the PBE results.

5.1.4. Summary. Statistics for the bond lengths, harmonic frequencies, and dissociation energies of the dimers studied are given in Table 9.

The very nature of the interatomic force in these rare gas dimers makes them a challenging case for DFT methods. The recent study of Tao and Perdew,[36] as well as that of Zhao and Truhlar,[50] have demonstrated that, although geometries accurate to around 10% are within the reach of many functionals, calculating accurate dissociation energies is considerably more problematic. To assist with the analysis of the small values involved, mean percentage errors for both the dissociation energies and the harmonic frequencies are given in Table 10.

The DFT-D methods employed in this study yield excellent geometries and reasonable harmonic frequencies but struggle to provide accurate dissociation energies. For the frequencies, B97-D stands head and shoulders above every other functional, and for dissociation energies HISS-D comfortably outperforms all of the other methods. For both properties, HSE-D does extremely well in 6 out of 10 cases, but its performance is marred by its inaccurate results for the neon-containing dimers. It is worth mentioning that PBE performs reasonably well for each property, but it is clear that the best performers overall are B97-D and HISS-D, with the consistently accurate energetics of the latter especially pleasing.

5.2. Solids. *5.2.1. Lattice Constants.* Table 6 lists the calculated lattice constants for the fcc structures of solid neon, argon, krypton, and xenon. The lattice constants are very well reproduced by all DFT-D methods, as well as the uncorrected M06-L functional, and although there is little to choose between them, B97-D and HISS-D are the most accurate by a small margin.

5.2.2. Sublimation Energies. Shown in Table 7, the sublimation energies of the rare gas solids appear to provide a significant challenge for DFT methods. Of the functionals to which an empirical correction has not been added, only M06-L reproduces the trend of increasing ΔH_{sub} down the group. However, it significantly overestimates the values for neon, argon, and krypton. HISS is the only functional to yield an accurate value for neon, but given its performance for the other solids this may be purely fortuitous. All of the DFT-D methods correctly reproduce the periodic trend, but all hugely overestimate the sublimation energy of solid neon. Rościszewski et al.[56] and Acocella et al.[57] have performed detailed additivity studies of the sublimation energy of solid neon, and both groups find that the zero-point energy contributes approximately 30% of the experimental value (around 600 J/mol), but this is not sufficiently large to explain the discrepancies observed here. For argon, krypton, and xenon, HSE-D performs very well, with a mean unsigned error of only 274 J/mol (0.07 kcal/mol) for these three solids, or 1290 J/mol (0.31 kcal/mol) if neon is included. Note that, as for the thermochemistry of the dimers, the energies involved here are extremely small. HISS-D yields a value slightly closer to experiment for neon and performs well for both krypton and xenon, but underestimates the sublimation energy of argon.

5.2.3. Band Gaps. The band gaps of each solid, calculated as minimum direct band energy differences at the optimized geometry for each functional, are given in Table 8. HSE performs very well, but by contrast the HISS functional is the least accurate of those tested. This is surprising given that HISS has been shown to be only slightly less accurate that HSE for this property,[10] but it is worth noting that these solids have unusually large experimental band gaps and that all of the functionals give qualitatively correct results. The empirical dispersion correction results in a slight improvement for HISS but degrades the results for all other functionals. Although the DFT-D correction has no direct bearing on electronic structure, it is disappointing that the general improvement in geometry observed with the correction does not translate into more accurate band gaps. TPSS performs surprisingly well, especially for the lighter elements; likewise, the statistics for M06-L are skewed somewhat by

Table 11. Errors in the Lattice Constants, Sublimation Energies, and Band Gaps of the Rare Gas Solids Studied

	a (Å)		ΔH_{sub} (J/mol)		band gap (eV)	
	MSE	MUE	MSE	MUE	MSE	MUE
HISS	0.5088	0.5121	−7658	7714	3.42	3.42
HSE	0.4485	0.4921	−6569	7082	0.22	1.02
M06-L	−0.1327	0.1327	3940	3940	−1.09	1.59
PBE	0.4565	0.5131	−5869	7043	−1.79	1.79
TPSS	0.8917	0.8917	−7314	7764	−0.62	0.95
B97-D	0.0630	0.1070	4389	4389	−1.75	1.75
BLYP-D	−0.1634	0.1634	2964	3308	−3.18	3.18
HISS-D	−0.0650	0.0989	−6	1567	2.31	2.52
HSE-D	−0.0542	0.1064	879	1290	−0.75	1.34
PBE-D	−0.1215	0.1215	4818	4818	−2.87	2.87
TPSS-D	−0.0673	0.1295	4497	4497	−2.01	2.01

the larger errors for krypton and, especially, xenon. While more extensive testing is warranted, this indicates that the changes in geometry due to adding the correction are insufficient to alter the band gap, and thus a functional that is adequate for band gaps will be adequate after correction.

5.2.4. Summary. Statistics for the lattice constants, sublimation energies, and band gaps of the solids studied can be found in Table 11.

Generally, traditional functionals perform poorly for these systems. Lattice constants are generally overestimated, and sublimation energies are wildly inaccurate. M06-L is the best of the uncorrected functionals for both properties, yielding reasonable lattice constants and qualitatively correct sublimation energies and band gaps.

The empirical dispersion correction makes a substantial difference to the computed lattice constants and sublimation energies, although all DFT-D methods except B97-D underestimate the lattice constant for neon. This may, in part, explain the hugely overestimated DFT-D sublimation energies for neon, although the value of ΔH_{sub} obtained with B97-D is also poor. Aside from neon, the accuracy of the HSE-D and HISS-D sublimation energies is excellent. As expected, HSE yields accurate band gaps for all of these solids, with TPSS and M06-L also doing well. The empirical correction is purely a function of nuclear geometry and not electronic structure, however, so it can have only a limited effect on the band gap.

As with the dimers, no functional is the best performer for every property, but we would argue that, overall, HSE-D is the most consistent. For geometries and energetics, HISS-D gives good results but is somewhat less accurate for band gaps.

6. Conclusions

The screened hybrid functionals HSE and HISS have been extended with the addition of an empirical dispersion correction. Within Grimme's DFT-D framework, we find that setting the global scaling parameter s_6 to 0.55 for both functionals gives the best results. The HSE-D and HISS-D functionals have been evaluated through calculations on a number of rare gas dimers and solids. Both were found to perform very well overall, with HISS-D performing slightly better for the dimers and HSE-D having the edge for the solids, although if the four neon dimers are excluded from the statistics, HSE-D is also the best functional for the molecular calculations. Given the differences between the complexes used in the parametrization and the molecules and solids investigated in this work, we suggest that both HISS-D and HSE-D should be applicable to both molecular and extended systems in which exact exchange and dispersion play an important part.

Acknowledgment. We are grateful to Dr. Petr Jurečka for providing geometries for the S22 test set and for helpful discussions about DFT-D parametrization, and Dr. Tom Henderson for his comments on the manuscript. This work was supported by E.N.B.'s Texas A&M University at Qatar startup funds and NPRP 08-431-1-076 from the Qatar National Research Fund.

References

(1) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

(2) (a) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207–8215. (b) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2006**, *124*, 219906.

(3) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 224106.

(4) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *121*, 1187–1192.

(5) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274–7280.

(6) Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. *J. Chem. Phys.* **2005**, *123*, 174101.

(7) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2009**, *11*, 443–454.

(8) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.

(9) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(10) Henderson, T. M.; Izmaylov, A. F.; Scuseria, G. E.; Savin, A. *J. Chem. Phys.* **2007**, *127*, 221103.

(11) Henderson, T. M.; Izmaylov, A. F.; Scuseria, G. E.; Savin, A. *J. Chem. Theory Comput.* **2008**, *4*, 1254–1262.

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

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

(14) Dobson, J. F.; White, A.; Rubio, A. *Phys. Rev. Lett.* **2006**, *96*, 073201−4.

(15) Ortmann, F.; Bechstedt, F.; Schmidt, W. G. *Phys. Rev. B* **2006**, *73*, 205101.

(16) Kerber, T.; Sierka, M.; Sauer, J. *J. Comput. Chem.* **2008**, *29*, 2088–2097.

(17) Civalleri, B.; Zicovich-Wilson, C. M.; Valenzano, L.; Ugliengo, P. *CrystEngComm* **2008**, *10*, 405–410.

(18) Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. *J. Phys. Chem. A* **2008**, *112*, 12530–12542.

(19) Savin, A. In *Recent Developments and Applications of Modern Density Functional Theory*; Seminario, J. M., Ed.; Elsevier: Amsterdam, 1996; Chapter 9, pp 327−357.

(20) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.

(21) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

(22) Zecca, L.; Gori-Giorgi, P.; Moroni, S.; Bachelet, G. B. *Phys. Rev. B* **2004**, *70*, 205127.

(23) Perdew, J. P.; Emzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(24) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(25) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(26) Kudin, K. N.; Scuseria, G. E.; Schlegel, H. B. *J. Chem. Phys.* **2001**, *114*, 2919–2923.

(27) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.

(28) 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.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian Development Version*, revision H.01; Gaussian, Inc.: Wallingford, CT, 2009.

(29) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(30) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(31) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **1999**, *110*, 7667–7676.

(32) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.

(33) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113–11123.

(34) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.

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

(36) Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2005**, *122*, 114102–7.

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

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

(39) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.

(40) (a) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865. (b) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396 .

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

(42) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101–18.

(43) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2009**, *130*, 074103.

(44) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(45) Ogilvie, J. F.; Wang, F. Y. H. *J. Mol. Struct.* **1992**, *273*, 277–290.

(46) Ogilvie, J. F.; Wang, F. Y. H. *J. Mol. Struct.* **1993**, *291*, 313–322.

(47) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **2003**, *118*, 4976–4983.

(48) Gerber, I. C.; Ángyán, J. G. *J. Chem. Phys.* **2007**, *126*, 044103.

(49) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I. *J. Phys. Chem. A* **2005**, *109*, 11015–11021.

(50) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 5121–5129.

(51) Aziz, R. A.; Slaman, M. *J. Chem. Phys.* **1989**, *130*, 187–194.

(52) Aziz, R. A. *J. Chem. Phys.* **1993**, *99*, 4518–4525.

(53) Steinmann, S. N.; Csonka, G.; Corminboeuf, C. *J. Chem. Theory Comput.* **2009**, *5*, 2950–2958.

(54) (a) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104−6. (b) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108–6.

(55) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; DiLabio, G. A. *J. Chem. Phys.* **2009**, *131*, 034111−7.

(56) Rościszewski, K.; Paulus, B.; Fulde, P.; Stoll, H. *Phys. Rev. B* **2000**, *62*, 5482.

(57) Acocella, D.; Horton, G. K.; Cowley, E. R. *Phys. Rev. B* **2000**, *61*, 8753.

(58) Batchelder, D. N.; Losee, D. L.; Simmons, R. O. *Phys. Rev.* **1967**, *162*, 767.

(59) Peterson, O. G.; Batchelder, D. N.; Simmons, R. O. *Phys. Rev.* **1966**, *150*, 703.

(60) Losee, D. L.; Simmons, R. O. *Phys. Rev.* **1968**, *172*, 944.

(61) Sears, D. R.; Klug, H. P. *J. Chem. Phys.* **1962**, *37*, 3002–3006.

(62) McConville, G. T. *J. Chem. Phys.* **1974**, *60*, 4093–4093.

(63) Schwalbe, L. A.; Crawford, R. K.; Chen, H. H.; Aziz, R. A. *J. Chem. Phys.* **1977**, *66*, 4493–4502.

(64) Magyar, R. J.; Fleszar, A.; Gross, E. K. U. *Phys. Rev. B* **2004**, *69*, 045111.

CT900536N

JCTC Journal of Chemical Theory and Computation

Formation of Interconnected Aggregates in Aqueous Dicationic Ionic Liquid Solutions

B. L. Bhargava*,†,‡ and Michael L. Klein†

Institute for Computational Molecular Science and Department of Chemistry, Temple University, 1900 N. 12th Street, Philadelphia, Pennsylvania 19122, and The Laboratory for Research on the Structure of Matter, University of Pennsylvania, 3231 Walnut Street, Philadelphia, Pennsylvania 19104-6202

Received December 15, 2009

Abstract: The structure and organization in an aqueous solution of a gemini surfactant, the dicationic ionic liquid 1,3-bis(3-decylimidazolium-1-yl) propane bromide, and its vapor−liquid interface have been studied using molecular dynamics simulations at room temperature. Starting from a uniform distribution of cations, the system is found to spontaneously evolve forming cross-linked cationic micellar aggregates. Alkyl tails are typically found buried inside the aggregates to minimize their unfavorable interactions with water, whereas the polar head groups are present at the micellar surfaces, exposed to water. Anions are found throughout the solution and are not strongly bound to the cations. Cationic micellar aggregates exhibit an interesting behavior: interconnection mediated by head groups, a phenomenon which is not observed in monocationic ionic liquid solutions. The structure of the vapor−liquid interface of the solution, the structure of the micellar aggregates, and the distribution of counterions are also discussed.

1. Introduction

Room temperature ionic liquids (RTILs) have interesting properties and potential applications.[1−3] Experimental efforts[4−7] and computational studies[8−12] have added to the knowledge on this new class of materials. Binary mixtures are known to possess properties not exhibited by either of the components. The area of applicability of ILs can be enormously increased by mixing them with other compounds and tailoring the mixture for a specific application. Several studies have been carried out in this direction, by mixing ILs with carbon dioxide,[13−15] hydrogen fluoride,[4] water,[16−18] and other compounds.

Aqueous solutions of ionic liquids have been studied extensively using experiments[19−22] and computational techniques,[23,24,16,25,26] the majority of them involving ILs based on imidazolium headgroups. These studies have aided in determining the structure of ILs in solutions. The chain length on the headgroup plays a major role in determining the structure of solutes in the solution. ILs with a small chain on the cation headgroup (≤ 4) remain as monomeric ions in solutions, whereas those with intermediate length chains ($n = 6-8$) form small clusters of cations. When the substituent on the headgroup is long alkyl chain ($n \geq 10$), the cations form micellar aggregates with the alkyl tails buried inside and the head groups lying at the surface.[18,26] Due to the inherent amphiphilic nature of these compounds, they can be used as surfactants.

Geminal dicationic ILs consist of a doubly charged cation that is composed of two singly charged cations linked by an alkyl chain (also called a spacer). A singly charged anion is associated with each of the charged parts of the cation in the crystalline phase. Such ILs with more than one polar and nonpolar region have been synthesized and characterized[27−31] and applied in chemical reactions[32,33] recently. Dicationic and tricationic ILs expand the horizon of applications for ionic liquids.[34−36] Crystal structures of some compounds belonging to this class have also been determined by Anderson et al.[27] ILs are also known to aggregate at very small concentrations and hence are useful as surfactants. A systematic study on monocationic and dicationic imidazolium bromide ILs with a tetradecyl chain has shown that the

* To whom correspondence should be addressed. Phone: +1-215-204-4217. Fax: +1-215-204-2257. E-mail: bhargav@sas.upenn.edu.

† Temple University.

‡ University of Pennsylvania.

10.1021/ct900674t
Published on Web 02/08/2010

dicationic compound is thermally more stable and also possesses a significantly lower CMC (about 200 times less).[28]

Dicationic ILs can be used as solvents and lubricants at high temperatures.[37] They also find applications in analytical chemistry, particularly in electrospray ionization mass spectrometry and in detecting small quantities of anions via gas phase ion association[38] and as the stationary phase in gas chromatography columns.[39] These group of ILs have received very little attention of researchers compared to the monocationic ILs. There are very few experimental studies on these compounds. Computational studies on dicationic ILs are limited to the electronic structure calculations in the gas phase[40] and semiempirical molecular modeling of interactions between ILs and a hydroxylated silicon surface.[41] To date, there seems to have been no efforts to study the microscopic structure of these materials in bulk using computational methods, such as molecular dynamics (MD). Accordingly, the present work looks into the structure and organization of a dicationic IL 1,3-bis(3-decylimidazolium-1-yl) propane bromide (two 1-*n*-decylimidazolium units linked by a $-(C_3H_6)-$ group) in bulk and the vapor/liquid interface of its aqueous solution. Details of the MD simulations are provided in the next section, which is followed by the results obtained and a discussion of these results. We end with the conclusions derived from the computational studies.

2. Methodology and Simulation Details

Classical MD simulations have been carried out on an aqueous solution of 1,3-bis(3-decylimidazolium-1-yl) propane bromide using the LAMMPS code[42] and the all-atom force field model developed by Pádua and co-workers.[43] [In the dicationic IL studied, the two imidazolium rings are connected to each other through a spacer, compared to the single imidazolium ring for which the model has been developed.[43] The residual charge (due to the replacement of the third hydrogen atoms of 3-methyl groups by a methylene group) has been distributed to the hydrogen atoms (+0.125*e* each) of the central methylene group of the spacer.] The simple point charge (SPC) model[44] has been used for water molecules.

A dicationic IL monomer was replicated in three dimensions and water molecules were added randomly to generate the starting configuration. The system was initially simulated in the isothermal−isobaric (constant NPT) ensemble at 1 atm pressure to fix the density. Equilibration at 1 atm pressure leads to a value of 76.86 Å for the edge length of the cubic box. Subsequent MD runs were performed in the canonical ensemble (constant NVT) with a box length derived from constant NPT simulations. The simulated system consisted of 125 cation entities, each with two units of positive charge, 250 bromide ions, and 11 317 water molecules. Three dimensional periodic boundary conditions were used to simulate the bulk behavior. A distance of 13 Å was chosen as the cutoff distance for computing nonbonded interactions. The potential was not shifted at the cutoff. Long range electrostatic interactions[45] were handled using the particle−particle particle-mesh solver (PPPM) with an accuracy of 1 part in 10^5. The equations of motion were integrated with a time step of 0.5 fs using the velocity Verlet algorithm, and SHAKE[46] was used to constrain the stretching and bending

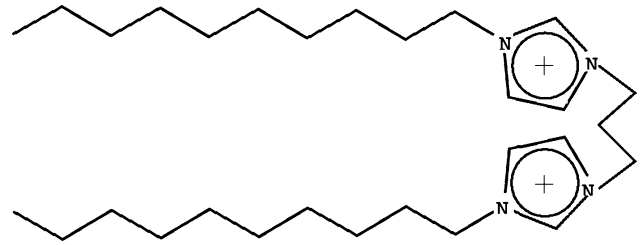


Figure 1. Schematic representation of the cation, 1,3-bis(3-decylimidazolium-1-yl) propane.

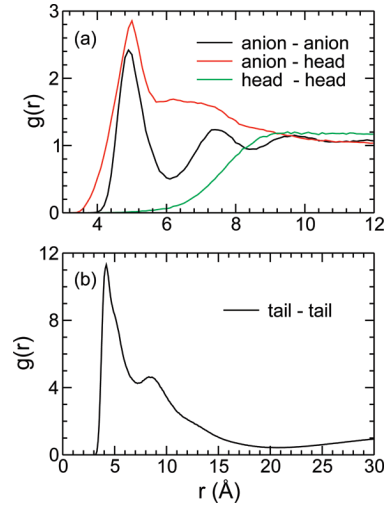


Figure 2. Radial distribution functions (a) between the headgroup and the anion (b) tail groups around themselves.

interactions of water molecules. A Nosé-Hoover thermostat and barostat with time constants of 1000.0 and 500.0 fs were used to control the temperature and pressure of the system, respectively. A trajectory of 40 ns was generated, out of which the final 30 ns was saved for analysis. Positions of the atoms were stored every 4 ps.

A well equilibrated system was placed at the center of a tetragonal box with sides of 76.86, 76.86, and 120.0 Å, to simulate a liquid/vacuum interface. Since the periodic boundary conditions were used on this super cell, the system represented an infinite number of thin films of aqueous IL solution separated by a vacuum. All other conditions remained the same as in the bulk simulations. The system was equilibrated for 10 ns and data for 30 more nanoseconds were saved for analysis.

The visualization software VMD[47] was used to render images of the simulations. A schematic of $[C_3(C_{10}Im)_2]^{2+}$ has been provided as Figure 1 to aid the discussion.

3. Results and Discussion

3.1. Radial Distribution Functions. The average distribution of different types of atoms, without the details of the orientation, can be obtained using pairwise radial distribution functions (RDFs). Different intermolecular RDFs are presented in Figure 2. Data are averaged over the last 10 ns of the trajectory. The center of the imidazolium ring refers to headgroup, and the terminal carbon atom of the decyl chain is referred to as a tail group in the figure. A binwidth of 0.1 Å is used in the calculation of RDFs. From Figure 2a, we

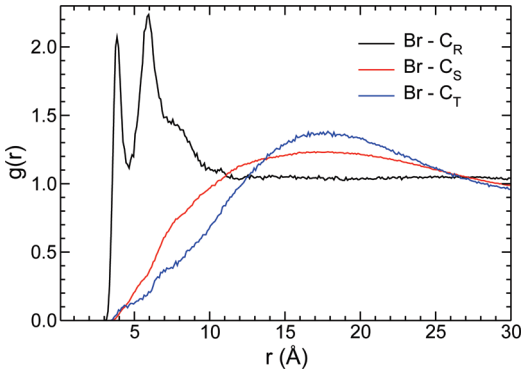


Figure 3. Radial distribution functions of C_R, C_S, and C_T around anions, where C_R's are the carbon atoms present between the two nitrogen atoms of the imidazolium ring, C_S's are the third to ninth carbon atoms of the decyl chain counting from the ring, and C_T's are the terminal carbon atoms of decyl chains.

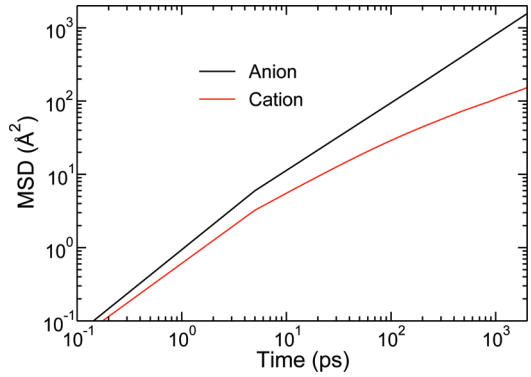


Figure 4. Mean squared displacements of the anion and cation. The data are averaged over the last 10 ns of the trajectory.

can see that the anion is very well organized around the cation. The anion−anion RDF shows a first maximum at 4.9 Å and has a coordination number of 0.43 up to the first minimum at 6.1 Å. The RDF also shows several noticeable maxima before the function attains its isotropic value of unity. The first peak in this function arises from the near neighbor bromide ions, whereas the subsequent peaks arises from the anions which are separated by a single water molecule, two water molecules, and so on. The anion−head RDF shows a sharp peak at 5 Å with an amplitude of 2.8, followed by a broad hump between 5.7 and 8.0 Å. The coordination number up to 8.0 Å is 1.8, which is the average number of head groups found within a distance of 8.0 Å from a particular anion. The RDF does not show any further features beyond 12 Å. The head−head radial distribution function shows a broad peak around 10 Å with an amplitude of 1.2. A broader second peak has been observed at around 24 Å (data not shown). The reason for the appearance of this peak will be discussed later in this section. Figure 2b shows the intermolecular radial distribution between the terminal carbon atoms of the decyl chain. The RDF peaks at 4.2 Å with an amplitude of 11.3, which suggests the extent of organization of decyl chains in the solution. The distribution is similar to that found in monocationic ILs with a long alkyl chain substituted on the ring.[26] The formation of aggregates is evident from the distribution of tail groups around themselves.

The radial distribution of atoms C_R (carbon atom between the two nitrogen atoms in the ring), C_S (third to ninth carbon atoms of decyl chain counting from the ring), and C_T (terminal carbon of decyl chain) around anions are shown in Figure 3. The RDF of C_R around the anion exhibits two sharp peaks at 3.95 and 5.95 Å. The first peak arises from the interaction of anions with the hydrogen attached to C_R. This hydrogen atom has been found to be more acidic compared to the other two hydrogen atoms on the imidazolium ring and form hydrogen bonds with the anion in pure ionic liquids and in IL solutions.[48,26] The second peak arises from the interaction with a second C_R atom present on the connected ring. It is interesting to note that the peaks for the C_S and C_T RDFs around the anion are broad and are present at around 18 Å, suggesting negligible interaction between the alkyl tail and anions. The peak position of the anion−C_T RDF provides an estimate of the median distance of anions from the terminal carbon atom of the decyl chain.

3.2. Hydrogen-Bonded Interactions. Hydrogen bonds (H-bonds) formed in the solution have been determined on the basis of geometric criteria. A hydrogen bond is said to have formed between the atoms if the distance between them is less than the sum of their van der Waals radii and they satisfy the linearity condition. We have chosen the H-bonds with a distance cutoff of 2.7 Å for bonds involving oxygen and 3.0 Å for bonds involving the bromide ion and an angle cutoff of 160° (spread < 20°) as strong H-bonds. We have characterized weak hydrogen bonds by adding 0.3 Å to the distance cutoff and choosing 140° as an angle cutoff. In the aqueous dicationic IL solutions, it has been observed that only 3% of the cations are bound to the anions with strong H-bonds. Including the weak H-bonds, this number increases to around 27%. Cations are also found to be H-bonded to water molecules. One in three cations are strongly bound to water via H-bonded interactions, and if we include the weak interactions it amounts to an average of 2.75 H-bonds per cation. Bromide ions are known to form strong H-bonds with water.[16] Each bromide ion is strongly bound to around 4 water molecules and weakly bound to two more water molecules. Hydrogen atoms of the imidazolium ring show varying degrees of association through H-bonds. While the acidic hydrogen atom (attached to the carbon atom between two nitrogen atoms of the ring) shows the highest affinity to forming H-bonds, the one bonded to the carbon atom linked with the nitrogen atom with a decyl chain shows the least affinity for H-bonds. This can be attributed to the stearic hindrance from the decyl chain.

3.3. Diffusion of Ions. Mean squared displacements (MSD) of the anion and cation are shown in Figure 4. The central carbon atom of the spacer propyl group present between the two imidazolium rings is considered as the cation position. The initial ballistic motion of the ions can be observed. From the figure, we can see that the MSD at 2 ns is around 1600 and 150 Å^2, respectively, for the anion and cation; i.e., the MSD of the anions is an order of magnitude higher than that of the cations. On average, anions are displaced by around 40 Å, whereas cations are displaced by a little above 12 Å. The ions diffuse considerably, and hence we can assume that the phase space is sampled adequately to give proper averages for the properties discussed. Anions

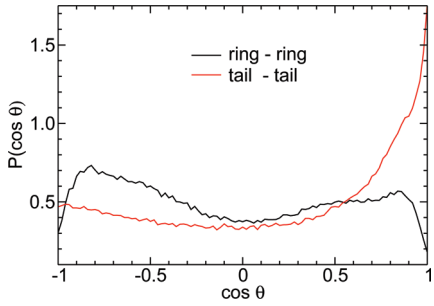


Figure 5. Orientation of polar and nonpolar groups of a cation.

are not strongly bound to cations, which is evident from the higher diffusion rate observed in the case of anions, compared to the cations, even though anions are likely to be found near the acidic hydrogen of the imidazolium ring. The reason for the slow diffusion of cations disproportionate to their mass is discussed later in this section.

3.4. Intramolecular Structure of Cations. The presence of two polar and two nonpolar groups in the cation can lead to a complex intramolecular structure. Both hydrophobic tails of the cation can interact with each other, or they may interact with the hydrophobic tails of another cation. The orientation of the two rings with reference to each other is also influenced by the conformation of the spacer. In Figure 5, the orientation probabilities for the rings and tails are shown: The *x* axis presents the cosine of the angle between two rings (measured as the angle between the normals to the planar rings) and two decyl tails (measured as the angle between the vectors connecting the first and last carbon atom of the decyl tails) belonging to the same cation, and the *y* axis presents the probability. It is clear from the figure that the tails belonging to a cation are likely to be oriented parallel to each other. However, the probabilities for nonparallel orientations are non-negligible. It can also be noticed that there is a higher probability of finding imidazolium rings of a cation oriented around 150° to each other. These are the most likely orientations, and not all the cations have the same intramolecular structure.

3.5. Formation of Aggregates. Cations are found to form aggregates in the aqueous solution in accordance with experimental findings.[28,29] This kind of behavior is also observed in monocationic IL solutions,[17,49,50,26] where the alkyl tails are found to be present at the core of the micellar aggregates. In the case of dicationic ILs, we observe similar behavior but with a difference. The decyl tails interact with each other, thus minimizing their exposure to water, initiating the aggregation process. It is also observed from a partitioning of the energy that the van der Waals interactions between the decyl tails are the main reason for aggregation. The imidazolium head groups surrounding these alkyl groups are thus shielding them from direct interaction with water while at the same time exposing them to favorable interactions. Anions interact with the head groups but are not bound to them strongly. Some of the cations act as linkers between micellar aggregates; i.e., they can be part of two aggregates with each of their decyl chains belonging to different micelles. This kind of interlinking (which is not observed in monocationic ILs[16]) between the aggregates makes the solution very viscous. The peak in the head−head RDF observed at around 24 Å arises from cations belonging to an aggregate and present on the diametrically opposite sides. Also, the observed slow diffusion of cations compared to the anions can be rationalized as due to the formation of interconnected aggregates.

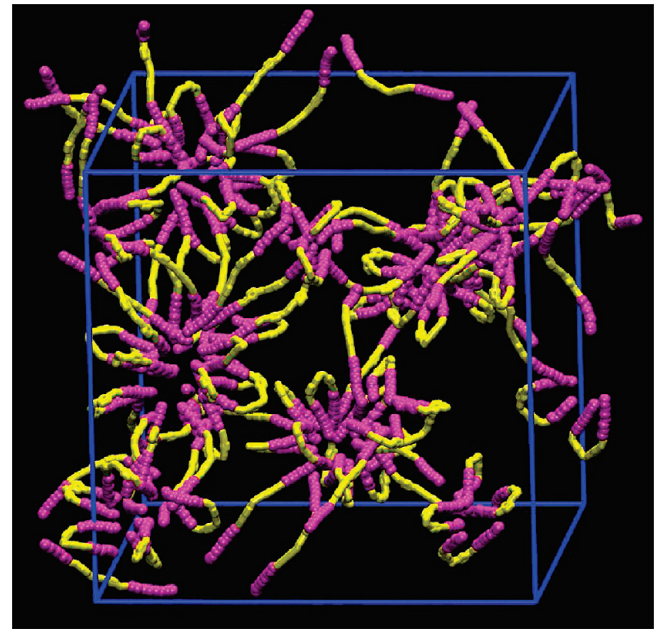


Figure 6. Aggregates of cations in an aqueous solution of $[C_3(C_{10}Im)_2]\cdot 2Br$. The positions are averaged over several frames. The yellow region represents the polar headgroup, while magenta spheres represent atoms belonging to the hydrophobic tail. Hydrogen atoms, anions, and water are not shown for the ease of visualization.

Figure 6 shows the observed aggregation of cations in aqueous solution. In the figure, atoms belonging to the headgroup of the cations are shown in a yellow color, while the atoms belonging to the hydrophobic tail region are shown as magenta spheres. Hydrogen atoms, bromide ions, and water molecules are not shown for clarity. The positions of the atoms are averaged over several frames near the end of the 40 ns simulation. We can notice the formation of near spherical micellar aggregates with the tail region in the core and head groups at the periphery. The links formed between two aggregates through the headgroup are visible in the figure. Anions show some preference to be present near the headgroup but are also found in the regions away from the imidazolium ring (data not shown).

3.6. Structure of an Aggregate. Formation of aggregates in solutions of amphiphilic cations is expected. However, the nature of the aggregates formed in this solution is rather interesting. Unlike those isolated near spherical aggregates formed in the monocationic IL solutions,[16] aggregates with a hydrophobic core and hydrophilic surface interconnected (cross-linked) to each other by the mediating polar region are found in dicationic IL solutions. From Figure 7, we can visualize the hydrophobic region (represented as magenta spheres) in the core of the near spherical aggregate interacting with each other via favorable dispersion interactions. Head groups (represented in yellow) surrounding the hydrophobic part interact with water while shielding the decyl chains from

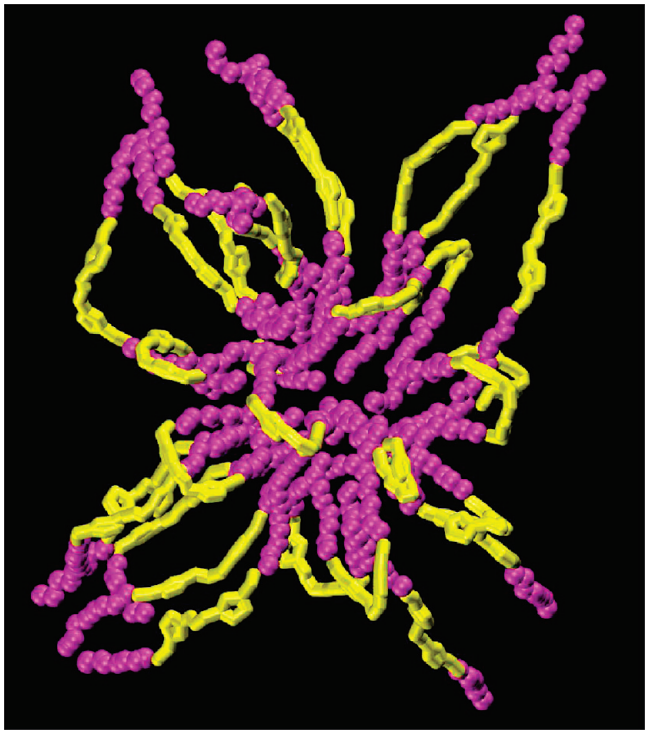


Figure 7. Structure of an aggregate in an aqueous solution of $[C_3(C_{10}Im)_2]\cdot 2Br$. Only the heavy atoms of cations belonging to the aggregate are shown in the figure.

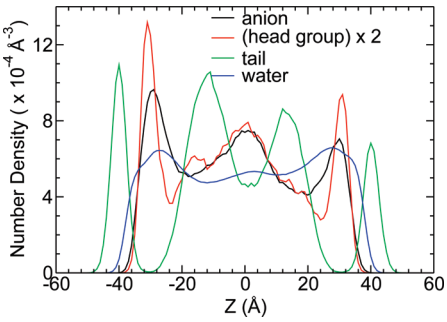


Figure 8. Number density profiles of anions, head groups, tail groups, and water in an aqueous solution of 1,3-bis(3-decylimidazolium-1-yl) propane bromide.

unfavorable exposure to water. Both decyl chains of some cations are involved in the formation of a single micellar aggregate, but others extend outside. In the figure, only one of the aggregate is completely shown, and the decyl chains extending out of the aggregate actually belong to a different aggregate (not shown in the figure).

3.7. Liquid−Vapor Interface. *3.7.1. Number Density.* The number density profiles of anions, head groups, tails, and water along the z axis, obtained from the vapor−liquid interface simulations, are shown in Figure 8. In these calculations, the headgroup is represented by the central carbon atom of the spacer ($-(CH_2)_3-$ group) located between the two imidazolium rings. The terminal carbon atoms of the decyl group represent tail groups. Since the ratio of cations to anions is 1:2 in the solution, we have accordingly multiplied the headgroup number density by two for comparison. Similarly the number density of water is normalized and presented. Notice the peak positions in the number densities of anion and headgroup are close to each other.

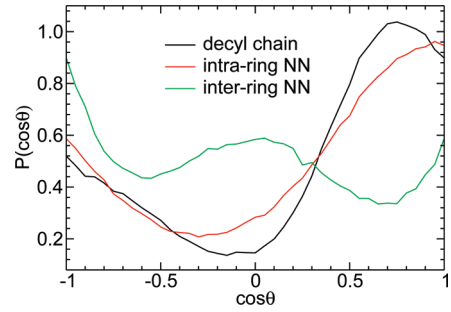


Figure 9. Orientation probability for vectors corresponding to different components of the cation along the z axis. Cos θ is the angle between the z axis and the vectors connecting the (black) first and last carbon atoms of the decyl chain, (red) two nitrogen atoms of an imidazolium ring, and (green) nitrogen atoms connected by the spacer.

However, the peak corresponding to cation is sharp near the interface due to the more organized orientation of cations near the interface. Moreover, the peaks are equally broad for both headgroup and anions in the bulk region of the solution. The reduction in the number density of water in the bulk region can be attributed to the formation of cation aggregates. The number density profile of the tail shows some important characteristics. The tail region is protruding out from the solution interface, which is evident from the profiles of the tail and water. At least some part of the decyl group is completely protruding out from the surface. We can see that the number density of tail groups is highest at the interface and suddenly drops to almost zero. We find two additional peaks in the tail number density in the bulk region of the solution. These peaks, which arise due to the presence of aggregates in the bulk, do not provide much insight, as the data are averaged over the surface. Observing the density profiles of the tail and water, it is obvious that the maxima for water are present near the minima for the tail and vice versa. So the high concentration of tails at a given distance from the interface indicates the formation of aggregates in the solution and hence decreases the available region for the water molecules.

Note the asymmetry present in the density profiles. Asymmetry is generally induced during the initial stages and persists for an extended period of time in the MD simulations. Here, we are only looking at the qualitative behavior of the ions in the aqueous solution. To obtain statistically significant quantitative values, it is advisable to simulate independent configurations and/or bigger systems for longer times.

3.7.2. Orientation of Cation. The orientation probabilities of vectors defining different components of the cation along the z axis are shown in Figure 9. The probabilities are computed only for the cations present beyond a certain cutoff distance from the origin along the z axis. The cutoff distance of 24 Å was chosen on the basis of the number density profile of cations. The probabilities are shown for the vectors connecting the first and last carbon atoms of the decyl chain (black curve), two nitrogen atoms belonging to the same imidazolium ring (red), and nitrogen atoms of different rings connected by the spacer (green). The probability of finding the decyl chain oriented parallel to the surface is the least and is very low (<0.15) compared to the isotropic value of 0.5. The most probable orientation is the chain tilted to the

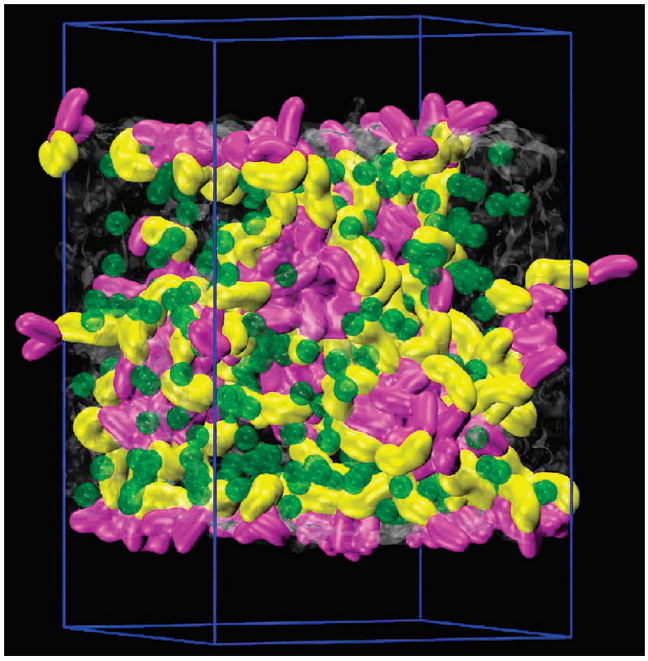


Figure 10. Volume map density of cations, anions, and water in the vapor−liquid interface of an aqueous solution of $[C_3(C_{10}Im)_2]\cdot 2Br$. Yellow represents the polar headgroup region, while magenta represents the hydrophobic tail group. Anion density is shown in green, and the water density is shown in transparent white.

interface normal. There is also a substantial probability of finding the chain oriented parallel to the interface normal. The vector connecting two nitrogen atoms of the ring is likely to be parallel to the interface normal. The orientation probability for the vector connecting nitrogen atoms separated by the spacer shows interesting behavior. We can see the peaks corresponding both to the parallel and perpendicular orientation of this vector with reference to the interface normal. The peak around zero corresponding to the perpendicular orientation arises from the cations that have both decyl chains at the interface parallel to each other, so that the spacer is perpendicular to these chains and hence to the z axis. The peak around 1 (or −1) corresponding to the parallel orientation along the z axis, arises from those cations with a decyl chain at the interface and the other inside the bulk region, belonging to an aggregate.

3.7.3. Organization at the Vapor−Liquid Interface. Observing the density profiles and orientation probability, it is evident that the decyl chains are present at the vapor−liquid interface of the dicationic IL solution partially protruding out of the water surface. These alkyl chains are likely to be tilted to the interface normal. A volume map density of the vapor−liquid interfacial system is presented in Figure 10. No ions were found in the vapor region of the interface. It can be observed from the figure that the segregation of polar and nonpolar regions of the cation that occurs is similar to that found in bulk MD simulations; namely, hydrophobic regions are surrounded by hydrophilic regions. The aggregates formed in the bulk region with decyl chains in the core are connected to each other via mediating head groups of the cations. Anions are found dispersed throughout the available region of the solution excluding the core of the aggregates.

4. Conclusions

The structure and organization in an aqueous solution of a dicationic ionic liquid and its liquid−vapor interface at a concentration higher than critical aggregation concentration (CAC) is studied using MD simulations at room temperature. Starting from a uniform distribution of cations, evolution of the system to form cationic aggregates has been observed. Anions are found to form a well-defined solvation shell of water via hydrogen-bonded interactions. Cationic head groups are also found to form H-bonds with water, though not as strongly as anions. Cations aggregate in such a way as to minimize the unfavorable interactions between the hydrophobic tail groups and water, while maximizing the favorable polar headgroup−water interactions. Thus, the core of the aggregate is devoid of water and anions, populated with only the decyl chains held together by van der Waals attraction. Partitioning of nonbonded interaction energies indicates the tail−tail interactions to be the major reason for the formation of aggregates.

Anions in the solution are present throughout the solution even though they are found to interact with the ring hydrogen atoms. The anions are not strongly bound to cations, which is evident from the diffusion of the ions. Unlike in monocationic aqueous IL solutions,[16] the aggregates in a dicationic IL solution are interconnected (cross-linked) to each other via mediating cations. Some of the cations are shared between two distinct aggregates with each of the head and tail groups belonging to one of them. The interconnection between the aggregates makes the solution highly viscous, which is also reflected in the diffusion of cations. The vapor−liquid interface of the aqueous solution is populated by the decyl chains that partially protrude out of the water surface. The alkyl tails are tilted to interface normal at a small angle. While some of the cations are entirely present in the interfacial region, others share their tails between the interface and the micellar aggregate present in the bulk. Interconnected aggregates are also found in the bulk region of the interfacial system.

In the present studies, MD simulations are performed on a very concentrated solution of 0.457 M, while the CAC values are very small, on the order of millimolar.[29] The current study only addresses the behavior of the solution at a higher concentration, and the structure of the aggregates at CMC may be different than that observed in these simulations. On the basis of the structure of the aggregates in the solution, IL−water mixtures with even higher concentrations of IL are likely to form bilayers of ILs. It will be interesting to see whether such a bilayer is made up of the folded or the extended form of the cation or a mixture of both, and also how they behave with changes in the IL concentration and the length of the spacer. These studies require simulations of large systems to be performed for a long duration of time and are feasible only through coarse-grained methods, with present day computational resources. Thus, it will be advisable to use coarse-grained MD simulations, to build on the present results and also to study the dicationic IL solutions at or near the critical aggregation concentrations.

Acknowledgment. We thank the Ras Al Khaimah Center for Advanced Materials for supporting this work through a research fellowship named for His Highness Sheikh Saqr Bin Mohammed Al Qasimi.

Supporting Information Available: Example LAMMPS input, parameter, and data files. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tong, J.; Liu, Q.; Xu, W.; Fang, D.; Yang, J. *J. Phys. Chem. B* **2008**, *112*, 4381–4386.

(2) Huddleston, J. G.; Visser, A. E.; Reichert, W. M.; Willauer, H. D.; Broker, G. A.; Rogers, R. D. *Green Chem.* **2001**, *3*, 156–164.

(3) Welton, T. *Chem. Rev.* **1999**, *99*, 2071–2084.

(4) Hagiwara, R.; Hirashige, T.; Tsuda, T.; Ito, Y. *J. Fluorine Chem.* **1999**, *99*, 1–3.

(5) Triolo, A.; Mandanici, A.; Russina, O.; Rodriguez-Mora, V.; Cutroni, M.; Hardacre, C.; Nieuwenhuyzen, M.; Bleif, H.; Keller, L.; Ramos, M. A. *J. Phys. Chem. B* **2006**, *110*, 21357–21364.

(6) Triolo, A.; Russina, O.; Bleif, H.; Di Cola, E. *J. Phys. Chem. B* **2007**, *111*, 4641–4644.

(7) Xiao, D.; Rajian, J. R.; Li, S.; Bartsch, R. A.; Quitevis, E. L. *J. Phys. Chem. B* **2006**, *110*, 16174–16178. Xiao, D.; Rajian, J. R.; Cady, A.; Li, S.; Bartsch, R. A.; Quitevis, E. L. *J. Phys. Chem. B* **2007**, *111*, 4669–4677.

(8) Hanke, C. G.; Price, S. L.; Lynden-Bell, R. M. *Mol. Phys.* **2001**, *99*, 801–809.

(9) Hu, Z.; Margulis, C. J. *Acc. Chem. Res.* **2007**, *40*, 1097–1105.

(10) Bhargava, B. L.; Balasubramanian, S. *J. Phys. Chem. B* **2008**, *112*, 7566–7573.

(11) Maginn, E. J. *J. Phys.: Condens. Matter* **2009**, *21*, 373101.

(12) Annapureddy, H. V. R.; Hu, Z.; Xia, J.; Margulis, C. J. *J. Phys. Chem. B* **2008**, *112*, 1770–1776.

(13) Huang, X.; Margulis, C. J.; Li, Y.; Berne, B. J. *J. Am. Chem. Soc.* **2005**, *127*, 17842–17851.

(14) Bhargava, B. L.; Krishna, A. C.; Balasubramanian, S. *AIChE J.* **2008**, *54*, 2971–2978.

(15) Shi, W.; Maginn, E. J. *J. Phys. Chem. B* **2008**, *112*, 2045–2055.

(16) Bhargava, B. L.; Klein, M. L. *J. Phys. Chem. A* **2009**, *113*, 1898–1904.

(17) Blesic, M.; Marques, M. H.; Plechkova, N. V.; Seddon, K. R.; Rebelo, L. P. N.; Lopes, A. *Green Chem.* **2007**, *9*, 481–490.

(18) Bhargava, B. L.; Klein, M. L. *Soft. Matter* **2009**, *5*, 3475–3480.

(19) Firestone, M. A.; Dzielawa, J. A.; Zapol, P.; Curtiss, L. A.; Seifert, S.; Dietz, M. L. *Langmuir* **2002**, *18*, 7258–7260.

(20) Zhang, H.; Liang, H.; Wang, J.; Li, K. *Z. Phys. Chem.* **2007**, *221*, 1061–1074.

(21) Zhao, Y.; Gao, S.; Wang, J.; Tang, J. *J. Phys. Chem. B* **2008**, *112*, 2031–2039.

(22) Blesic, M.; Lopes, A.; Melo, E.; Petrovski, Z.; Plechkova, N. V.; Lopes, J. N. C.; Seddon, K. R.; Rebelo, L. P. N. *J. Phys. Chem. B* **2008**, *112*, 8645–8650.

(23) Jiang, W.; Wang, Y.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4812–4818.

(24) Wang, Y.; Jiang, W.; Yan, T.; Voth, G. A. *Acc. Chem. Res.* **2007**, *40*, 1193–1199.

(25) Bhargava, B. L.; Klein, M. L. *Mol. Phys.* **2009**, *107*, 393–401.

(26) Bhargava, B. L.; Klein, M. L. *J. Phys. Chem. B* **2009**, *113*, 9499–9505.

(27) Anderson, J. L.; Ding, R.; Ellern, A.; Armstrong, D. W. *J. Am. Chem. Soc.* **2005**, *127*, 593–604.

(28) Ding, Y.; Zha, M.; Zhang, J.; Wang, S. *Colloids Surf., A* **2007**, *298*, 201–205.

(29) Baltazar, Q. Q.; Chandawalla, J.; Sawyer, K.; Anderson, J. L. *Colloids Surf., A* **2007**, *302*, 150–156.

(30) Liu, X.; Xiao, L.; Wu, H.; Chen, J.; Xia, C. *Helv. Chim. Acta* **2009**, *92*, 1014–1021.

(31) Lovelock, K. R. J.; Deyko, A.; Corfield, J.-A.; Gooden, P. N.; Licence, P.; Jones, R. G. *ChemPhysChem* **2009**, *10*, 337–340.

(32) Han, X.; Armstrong, D. W. *Org. Lett.* **2005**, *7*, 4205–4208.

(33) Liu, Q.; van Rantwijk, F.; Sheldon, R. A. *J. Chem. Technol. Biotechnol.* **2006**, *81*, 401–405.

(34) Zhang, Z. X.; Zhou, H. Y.; Yang, L.; Tachibana, K.; Kamijima, K.; Xu, J. *Electrochim. Acta* **2008**, *53*, 4833–4838.

(35) Zafer, C.; Ocakoglu, K.; Ozsoy, C.; Icli, S. *Electrochim. Acta* **2009**, *54*, 5709–5714.

(36) Li, X.; Bruce, D. W.; Shreeve, J. M. *J. Mater. Chem.* **2009**, *19*, 8232–8238.

(37) Jin, C.; Ye, C.; Phillips, B. S.; Zabinski, J. S.; Liu, X.; Liu, W.; Shreeve, J. M. *J. Mater. Chem.* **2006**, *16*, 1529–1535.

(38) Nachtigall, F. M.; Corilo, Y. E.; Cassol, C. C.; Ebeling, G.; Morgon, N. H.; Dupont, J.; Eberlin, M. N. *Angew. Chem., Int. Ed.* **2008**, *47*, 151–154.

(39) Han, X.; Armstrong, D. W. *Acc. Chem. Res.* **2007**, *40*, 1079–1086.

(40) Sun, H.; Zhang, D.; Liu, C.; Zhang, C. *THEOCHEM* **2009**, *900*, 37–43.

(41) Nooruddin, N. S.; Wahlbeck, P. G.; Carper, W. R. *Tribol. Lett.* **2009**, *36*, 147–156.

(42) Plimpton, S. J. *J. Comput. Phys.* **1995**, *117*, 1–19; [Online] http://lammps.sandia.gov.

(43) Lopes, J. N. C.; Deschamps, J.; Pádua, A. A. H. *J. Phys. Chem. B* **2004**, *108*, 2038–2047. *108*, 11250;Lopes, J. N. C.; Pádua, A. A. H. *J. Phys. Chem. B* **2006**, *110*, 19586–19592.

(44) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, 1981; p331−342.

(45) Allen, M. P.; Tildesley, D. J. *In Computer Simulation of Liquids*; Clarendon: Oxford, 1987.

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

(47) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(48) Del Popolo, M. G.; Lynden-Bell, R. M.; Kohanoff, J. *J. Phys. Chem. B* **2005**, *109*, 5895–5902.

(49) Inoue, T.; Dong, B.; Zheng, L. *J. Colloid Interface Sci.* **2007**, *307*, 578–581.

(50) Vanyúr, R.; Biczók, L.; Miskolczy, Z. *Colloids Surf., A* **2007**, *299*, 256–261.

CT900674T

JCTC Journal of Chemical Theory and Computation

Acidity of the Aqueous Rutile TiO_2(110) Surface from Density Functional Theory Based Molecular Dynamics

Jun Cheng and Michiel Sprik*

Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Received January 7, 2010

Abstract: The thermodynamics of protonation and deprotonation of the rutile TiO_2(110) water interface is studied using a combination of density functional theory based molecular dynamics (DFTMD) and free energy perturbation methods. Acidity constants are computed from the free energy for chaperone assisted insertion/removal of protons in fully atomistic periodic model systems treating the solid and solvent at the same level of theory. The pK_a values we find for the two active surface hydroxyl groups on TiO_2(110), the bridge OH (Ti_2OH^+), and terminal H_2O adsorbed on a 5-fold Ti site ($TiOH_2$) are -1 and 9, leading to a point of zero proton charge of 4, well within the computational error margin (2 pK_a units) from the experimental value (4.5−5.5). The computed intrinsic surface acidities have also been used to estimate the dissociation free energy of adsorbed water giving 0.6 eV, suggesting that water dissociation is unlikely on a perfect aqueous TiO_2(110) surface. For further analysis, we compare to the predictions of the MUltiSIte Complexation (MUSIC) and Solvation, Bond strength, and Electrostatic (SBE) models. The conclusion regarding the MUSIC model is that, while there is good agreement for the acidity of an adsorbed water molecule, the proton affinity of the bridging oxygen obtained in the DFTMD calculation is significantly lower (more than 5 pK_a units) than the MUSIC model value. Structural analysis shows that there are significant differences in hydrogen bonding, in particular to a bridging oxygen which is assumed to be stronger in the MUSIC model compared to what we find using DFTMD. Using DFTMD coordination numbers as input for the MUSIC model, however, led to a pK_a prediction which is inconsistent with the estimates obtained from the DFTMD free energy calculation.

1. Introduction

Metal oxides develop a sizable positive surface charge when immersed in water of sufficiently low pH. The origin of the excess charge is protonation of basic surface oxygens. Similarly, at high pH, deprotonation of adsorbed water molecules or hydroxyl groups builds up a negative surface charge. This charging process controls the sorption of ions and surface speciation and, hence, affects the chemical reactivity of the metal oxide surface.[1] The surface charge density at given pH is however not only determined by the proton affinity or acidity of surface groups. It also depends on surface composition and the electrostatic potential difference across electrical double layers and therefore on the structure of the electrical double layer. It is notoriously difficult to disentangle these three factors using only experimental data such as potentiometric titration curves and electrokinetic measurements. Modeling and prediction of "intrinsic" surface proton affinities has therefore played a crucial role in the understanding of surface protonation.[2−8]

This complexity also leaves surface protonation models a large degree of freedom. Indeed, two of the most developed and successful models used in the literature, the MUltiSIte Complexation (MUSIC) model[2,3,5] and the Solvation, Bond strength, and Electrostatic (SBE) model,[4,6] are capable of representing a large set of experimental data. Both models use bond valence as the key parameter determining proton affinity. Bond valence was introduced by Pauling to rationalize the structure of ionic crystals. To predict intrinsic pK_a, a

* To whom correspondence should be sent. E-mail: ms284@cam.ac.uk.

10.1021/ct100013q
Published on Web 02/12/2010

bond length dependent generalization due to Brown and Altermatt is used.[9] The precise form of the correlation between pK_a and bond valence is however different in the MUSIC and SBE models[3,6] (the more recent model of Bickmore et al.[7,8] employs yet another relation). Also, the type of experimental data used for the parametrization is not the same. Equilibrium constants for surface protonation in the MUSIC model are based on a linear correlation of bond valence with experimental acidities of (hydr)oxyacids in homogeneous solution,[3] while the SBE model fits directly to experimental surface acidities.[4] A further important distinction is the treatment of solvation, hydrogen bonding, and local structure of the surface. The SBE model maintains the classical single site-two pK_a concept considering only a single generic surface hydroxyl group.[1] Long range interactions with the solvent (and the oxide) are described by the coupling to a dielectric continuum.[4] The MUSIC model differentiates between surface sites, recognizing that the number of metal ions coordinating with surface oxygens is a key factor in the differences in their chemical behavior.[2,3] Short range specific hydrogen bonding is also accounted for by assigning a bond valence to hydrogen bonds to basic oxygens.[3] It is however this feature special to the MUSIC model that has been called recently into question[6−8] and will also be investigated in the present contribution.

In view of the variety in assumptions made in these models, a less empirical approach could be instructive even if only applied to a subset of typical systems. A significant step toward this goal was made by Rustad and co-workers in their studies of iron (hydr)oxide[10,11] and silica.[12] While their model is also parametrized using the pK_a of solution complexes, the molecular quantity correlated with these data is the proton binding energy of the corresponding gas-phase complex computed by molecular mechanics (MM) methods. Surface acidity constants are obtained by applying the MM Hamiltonian to an atomistic model of the gas-phase surface and substituting the calculated proton affinity in expression for pK_a. The advantage over the MUSIC and SBE approach is that the effect of the structural relaxation of surfaces and coupling between protonated sites can be studied in microscopic detail.[11] The importance of a realistic description of metal oxygen surface bonds is also stressed by Bickmore and co-workers.[7] The valence of surface bonds in their approach is calculated from detailed information on bond lengths as determined by full Density Functional Theory (DFT) modeling of periodic oxide vacuum interfaces. This method was applied in an investigation of the protonation of aluminum hydroxide(gibbsite) and silica.[7] A similar strategy was used by Machesky and co-workers to obtain the structural input for a MUSIC model estimation of surface protonation constants of TiO_2 (rutile).[13] The Ti−O bond lengths were estimated from a density functional theory based molecular dynamics (DFTMD) simulation of multilayer adsorption of water on TiO_2 surfaces. pK_a in a MUSIC model, however, also depends on coordination numbers (number of hydrogen bonds) which were computed using a classical MD simulation of solid−water interfaces. The force field model in the classical MD was optimized using DFT,[14,15] ensuring consistency between the two calculations.

A feature common to all methods mentioned above is reliance on an empirical linear free energy relationship of some kind for the description of the solvent effects on pK_a. Eliminating such a phenomenological relation requires either monitoring the surface protonation as it evolves in a MD simulation allowing for proton dissociation or the application of free energy sampling methods. Examples of the first approach are the classical MD studies of the charging of magnetite[16] and goethite[17] by Rustad and co-workers using dissociative water potentials. The present DFTMD investigation of the protonation of the rutile TiO_2(110) surface uses free energy perturbation methods. Proton affinities are computed directly as finite temperature free energy changes using our recently developed DFTMD method of reversible proton insertion.[18,19] The motivation for choosing the TiO_2/H_2O interface is because this system is well characterized by experiment[20−23] and has become a benchmark for the modeling of surface protonation and complexation.[2−5,13,21,24]

A related issue, which has caused considerable controversy in the computational surface science community, is whether water dissociates on the rutile TiO_2(110) surface.[25−35] Experiments of TiO_2 surfaces exposed to a low density water vapor seem to indicate that water cannot dissociate on the perfect surface except at defect sites (i.e., O vacancies).[36−38] The first DFT calculations came to the opposite conclusion (see ref 35 for a recent review). An important step was made by Lindan and co-workers,[26] who showed the importance of inter-adsorbate hydrogen bonding at higher coverages, thus distinguishing mono- and multilayer coverage from submonolayer systems for which the disagreement between theory and experiment is most pronounced. However, a clear agreement on the structure of a H_2O monolayer adsorbed on TiO_2 rutile (110) is also still lacking. DFT calculations on this system vacillate between molecular (associative)[25,29,30,33−35] or mixed associative and dissociative adsorption.[26−28,31,32] These conclusions are based on a comparison of adsorption energies computed from total energies of systems placed in vacuum. The differences per H_2O molecule are however often small and dependent on a variety of conditions (see section 3.2). In contrast, in the present approach, the free energy for water dissociation at the TiO_2/H_2O interface is computed by combining the acidities of surface groups taking into account the solvation of surface (hydr)oxide groups.

2. Theory and Methods

2.1. Surface Protonation Model and Point of Zero Charge. On rutile TiO_2(110) there are two types of surface sites capable of binding additional protons under normal pH conditions (see Figure 1): the hydroxylated 5-fold coordinated Ti ("titanol") groups

$$TiOH^- + H^+ \rightarrow TiOH_2,\ K_{H1} \quad (1)$$

and bridging oxygens

$$Ti_2O + H^+ \rightarrow Ti_2OH^+,\ K_{H2} \quad (2)$$

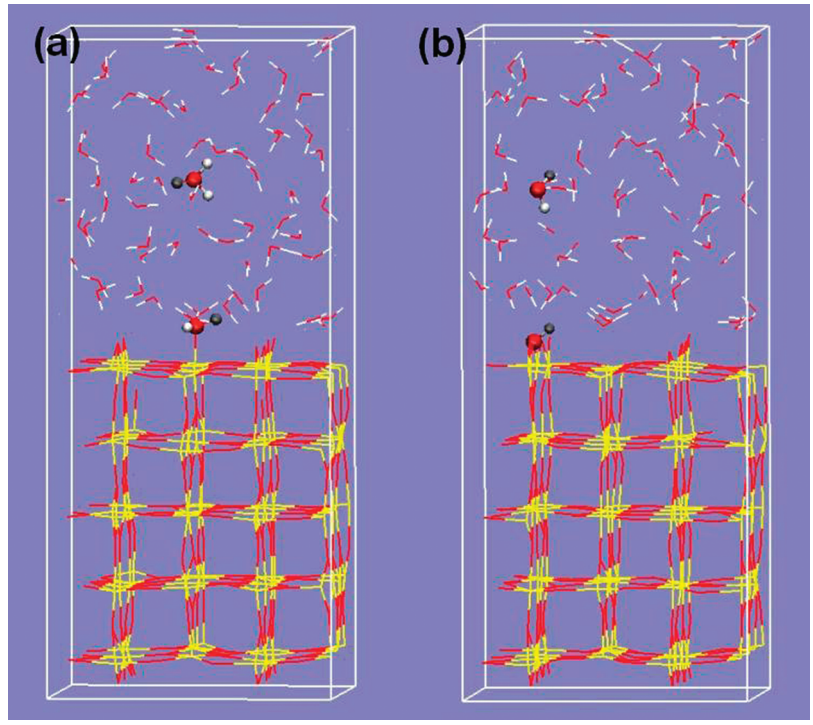


Figure 1. Molecular dynamics model system and schematic representation of the method for the computation of acidity constants of surface (hydr)oxide groups at the rutile TiO_2(110)/H_2O interface. The pictures show the full MD supercell in the (a) $TiOH_2$/$TiOH^-$ and (b) Ti_2OH^+/Ti_2O conformations. These systems have been set up as 5 O−Ti−O trilayers and 71 water molecules. 3D periodic boundary conditions are applied leading to an alternation of TiO_2 slabs and water layers. Ti, O, and H atoms are distinguished in yellow, red, and white, respectively. The molecules involving protonation/deprotonation reactions are highlighted by a ball and stick representation. The gray balls denote the inserted/annihilated protons. Switching one proton on and the other off in a and b simulates eqs 9 and 11, respectively.

where K_{H1} and K_{H2} are the corresponding protonation equilibrium constants. Following the notation of ref 2, the subscripts 1 and 2 to the equilibrium constants refer to the metal coordination of the active oxygen.

Charges in schemes 1 and 2 have been assigned according to the formal charge of an adsorbed OH_n group ($n = 0, 1, 2$) assuming that the OH_2 species (a water molecule) is neutral. The TiO_2 solid, represented in the DFTMD model by a finite slab, is therefore viewed as a large molecular unit (cluster) which is neutral when associatively hydrated by water molecules. This simple ionic picture, based on integer proton charge only, ignores contributions from the coordinated Ti ions to the surface charge. This effect is taken into account in the MUSIC model by adding in the fractional Pauling bond valence (+2/3) of the TiO bond. Reactions 1 and 2 are then written as[3,24]

$$TiOH^{-1/3} + H^+ \rightarrow TiOH^{+2/3} \quad (3)$$

$$Ti_2O^{-2/3} + H^+ \rightarrow Ti_2OH^{+1/3} \quad (4)$$

While this may be a more realistic model of surface charge, such a model is not needed in a DFTMD calculation of the corresponding protonation free energy. The simple "pseudo" molecular representation of eqs 1 and 2 therefore seems more appropriate in this context. Note that the charges in scheme 1 also differ from the charges in the single-site two-pK_a model of ref 4 in which the TiOH group is formally neutral with a positive conjugate acid $TiOH_2^+$. The argument for relating the point of zero net proton charge (PZC) to the acidities of the surface groups used in ref 4 however still applies in our two-site model. The PZC is derived by combining two successive protonation reactions to a reaction reversing the sign of the surface charge. In our case, these are reactions 1 and 2, which leads to

$$TiOH^- + Ti_2O + 2H^+ \rightarrow TiOH_2 + Ti_2OH^+,\ K_{PZC} \quad (5)$$

At the PZC equilibrium, surface concentrations of $TiOH^-$ and Ti_2OH^+ are equal. Moreover, the ratio of 5-fold coordinated Ti^{4+} ions to bridging oxygens on a 110 face is 1:1. Substituting in the chemical equilibrium equation of reaction 5, we find $[H^+]_{PZC} = (K_{PZC})^{-1/2} = (K_{H1}K_{H2})^{-1/2}$ or in terms of pH and pK_a units using pK_{an} = log K_{Hn}

$$PZC = \frac{1}{2}(pK_{a1} + pK_{a2}) \quad (6)$$

Reaction 2 can also be coupled with the reverse of reaction 1, giving

$$TiOH_2 + Ti_2O \rightarrow TiOH^- + Ti_2OH^+,\ K_d \quad (7)$$

This process can be interpreted as the dissociation of an adsorbed water molecule (leaving the total surface charge the same). Since $K_d = K_{H2}/K_{H1}$, the corresponding free energy change $\Delta A_{diss} = -k_BT \ln K_d$ is the difference in acidities of the two conjugate acids:

$$\Delta A_{diss} = 2.30 k_B T(pK_{a1} - pK_{a2}) \quad (8)$$

2.2. Calculation of Acidities. The proton affinities of the surface (hydr)oxide groups were computed using a combination of DFTMD simulation and free energy perturbation (FEP) methods.[18] The role of the MD is to sample a mapping potential consisting of a linear mixing of the Born−Oppenheimer energy surfaces of reactant and product states. Free energy changes are obtained by integrating ensemble averages of vertical energy gaps along the alchemical transformation path from reactant to product. This method effectively inserts acid protons or deletes them. A similar FEP approach has been used by our group for the DFTMD estimation of redox free energies of inorganic[39−42] and organic[19,43,44] ions or molecules in solution. The method for the computation of acidities is more involved as it is more difficult to add or remove protons from a condensed phase system than adding or removing electrons. Rather than eliminating the proton entirely from the system, its charge is switched off with a harmonic restraining potential holding it in place. This method can be regarded as a DFTMD implementation of the chaperone assisted methods for reversible insertion of protons which have been applied, for example, for computation of tautomerization free energies of organic molecules.[45] For a detailed description, we refer to our previous publications.[18,19] A brief summary is given in the Supporting Information.

The insertion of a single proton can be regarded as a half reaction.[19] The DFTMD model systems to which the protons are added are the usual periodically repeated supercells of molecular dynamics simulation. The reference (zero) of the electrostatic potential in such a setup is artificial and has no physical meaning.[19,46,47] Addition or removal of a single ion changes the net charge of the cell and the corresponding free energy can therefore not be identified with the absolute solvation free energy (even in the limit of large cell size). The discrepancies for a typical DFTMD model system are significant (3 eV or more).[19,42] This bias cancels when deprotonation of a species or group is carried out in conjunction with protonation of another species in the same cell, thus avoiding a change of net charge. However, the drawback of such a full reaction scheme compared to a half reaction scheme is that model systems must be considerably larger in order to minimize the interaction between charged acid or base species (recall that in order to reproduce standard conditions these interactions must be eliminated). Furthermore, because of the powerful screening properties of water, the effective interactions of ions with their periodic images and compensating background charge in MD models of homogeneous solution are surprisingly small[48] (see also ref 19). This is the reason why we preferred a half reaction scheme based on the insertion of single protons in our previous calculations of the acidity of aqueous species in homogeneous solution.[18,19] Provided the model systems have an excess of solvent, the sum of the free energies of a protonation and deprotonation half reaction is equally unaffected by a shift in the reference of the electrostatic potential because the reference of the electrostatic potential in a low concentration solution is determined by the solvent and is therefore the same in the two half reactions.

A solid liquid interface is a highly inhomogeneous system. Periodic boundary effects are likely to be different from homogeneous solutions, which is why we decided to use the more direct full reaction scheme in our first study of the TiO_2−water interface. This scheme is illustrated in Figure 1a. Deprotonation of $TiOH_2$ to $TiOH^-$ is coupled to a simultaneous protonation of H_2O in the liquid water part of the same inhomogeneous model system. The reverse of this reaction amounts to a proton transfer from a hydronium in solution to a $TiOH^-$ surface group:

$$TiOH^- + H_3O^+ \rightarrow TiOH_2 + H_2O \quad (9)$$

and is equivalent to the surface protonation reaction eq 1. The free energy change of reaction 9, which will be indicated by ΔA_1, is obtained from a coupling parameter integral of the corresponding vertical energy gap. Referring ΔA_1 to the pK_a of H_3O^+ (= −1.74) gives an estimate of the pK_a of $TiOH_2$:

$$pK_{a1} = pK_{a,H_3O^+} - \frac{1}{2.30 k_B T}\Delta A_1 \quad (10)$$

Similarly, the protonation of a Ti_2O group (reaction eq 2) must also be balanced by a reference deprotonation (Figure 1b):

$$Ti_2O + H_2O \rightarrow Ti_2OH^+ + OH^- \quad (11)$$

The free energy change of reaction 11 will be indicated by ΔA_2. The argument for deprotonating a H_2O molecule rather than a H_3O^+ is that the computation of the PZC and water dissociation constant requires that the reactant state be the same as the product state in reaction 9 (note that this also ensures that the molecular dynamics keep a zero net charge at all times). The pK_a of Ti_2OH^+ is found from the pK_b of Ti_2O. Adding the pK_a of H_2O (= 15.74) to (minus) the free energy change ΔA_2 of reaction 11 yields the expression

$$pK_{a2} = pK_{a,H_2O} - \frac{1}{2.30 k_B T}\Delta A_2 \quad (12)$$

The values used for pK_{a,H_3O^+} and pK_{a,H_2O} need some clarification. pK_{a,H_3O^+} is a constant and by definition equal to $-\log[\rho_{H_2O}/c^\circ] = -1.74$ where $\rho_{H_2O} = 55.5$ mol dm^{-3} is the ambient density of liquid water and $c^\circ = 1$ mol dm^{-3} is the standard concentration of solution chemistry. The reason why the reaction free energy ΔA_1 must be corrected by this term when converted to an acidity on the pK_a scale is that our method is based on the Brønsted picture of acidity in which acid dissociation is treated as a proton transfer to the solvent.[19] Accordingly, in the present application, reaction 1 is replaced by reaction 9. Similarly, the reference term in eq 12 is not pK_w, the dissociation constant of water, but $pK_{a,H_2O} = pK_w - pK_{a,H_3O^+} = 14.0 + 1.74$. To be strictly consistent, a DFTMD computed value for pK_w should be used. This calculation is under way in our group. In default of this result, the experimental value is used instead.

A further comment concerns the bias introduced by the restraining potentials. Let us first reiterate that these restraining potentials play an absolutely crucial role. There is first

of all the rather practical consideration that the dummy proton is invisible to its environment and, when not restrained, can wander off anywhere in the system. The restraining potentials keep the dummy close to where it was, avoiding the highly unstable configurations that might otherwise arise when the charge is switched back on.[18] However, there is also a more fundamental side to the application of restraints. We are interested in the acidity of a specific functional group. The proton is therefore removed from a group of this kind and must also be reattached to this group. This species specific insertion is directed by the restraining potentials. These potentials will however inevitably introduce a bias in the free energy for which we must correct. In ref 19, this question is approached by resolving the transfer of a proton from an acid AH to a H_2O molecule into a Born−Haber cycle involving the acid proton. The acid proton is transferred to the gas phase and then reinserted in solution attached to a water molecule. This fictitious process makes it easier to keep track of the various entropy and zero point motion contributions. The result is a thermochemical correction which is specified for the present system in the Supporting Information. This thermochemical correction for a half reaction can be substantial (amounting for a H_2O molecule to about 0.5 eV).[19] For full reactions (proton transfers), these corrections cancel to a large extent but not completely (see below).

2.3. Computational Implementation. The rutile TiO_2-(110) surface was modeled by periodic slabs of five O−Ti−O trilayers with lateral dimensions of a 4×2 surface cell. The slabs are separated by a space of 15 Å leading to an orthorhombic supercell cell of $11.9 \times 13.2 \times 30.8$ Å^3. Full 3D periodic boundary conditions are applied. To model the TiO_2/H_2O interface, the space between the TiO_2 slabs is fully filled with 71 water molecules so that there are two symmetric interfacial planes in each unit cell. The number of water molecules has been chosen to adjust the effective density of the water layer to the ambient density of water. A further consideration was to make the volume of water in a supercell approximately cubic in order to minimize confinement effects.

The gradient-corrected Perdew−Burke−Ernzerhof (GGA-PBE) functional[49] was used for all calculations. The DFTMD simulations use the Born−Oppenheimer method and have been carried out using the freely available program package CP2K/Quickstep.[50,51] The density functional implementation in Quickstep is based on a hybrid Gaussian plane wave (GPW) scheme. Orbitals are described by an atom-centered Gaussian-type basis set, while an auxiliary plane wave basis is used to re-expand the electron density.[52] The wave function optimization is performed using an orbital transformation minimizer, which avoids the traditional matrix diagonalization method and gives optimal convergence control.[53] Analytic Goedecker−Teter−Hutter (GTH) pseudopotentials[54,55] have been employed to represent the core electrons. The basis sets for the valence electrons ($2s^2 2p^4$ for O and $3s^2 3p^6 3d^2 4s^2$ for Ti) consist of short-ranged (less diffuse) double-ζ basis functions with one set of polarization functions (DZVP).[56] The plane wave basis for the electron density is cut off at 280 Ry. All our simulations only use the Γ point of the supercell for expansion of the orbitals. The convergence criterion for wave function optimization is set by a maximum electronic gradient of 3×10^{-7} and an energy difference tolerance between self-consistent field (SCF) cycles of 10^{-13}. We should admit that because of the very large system required to model a TiO_2/H_2O interface the present study uses smaller basis sets than our previous work (TZV2P).[18,19] This may be justified by two facts: (i) the currently used basis sets reproduce other DFT calculations of water adsorption energies with various configurations and coverages on TiO_2(110) with good accuracy[26,29] (see section 3.2), and (ii) dynamical effects in water are not very sensitive to the basis sets used.[57]

The time step for the MD simulation was 0.5 fs. NVT conditions were imposed by a Nose-Hoover thermostat with a target temperature of 330 K. All settings have been tested in previous work of our and other groups and have been proven to be sufficient to give a reasonable representation of structural and dynamical properties of liquid water at room temperature.[57] It should be noted that the elevated temperature of 330 K is chosen to avoid the glassy behavior of PBE liquid water on the 20 ps time scale observed for trajectories at lower temperatures.[57] In MD runs, 1−2 ps of equilibration period is followed by ∼5 ps of production period. This duration is adequate to obtain sufficiently accurate estimates of the vertical energy gap ΔE of reactions 9 and 11. The corresponding free energies ΔA_1 and ΔA_2 were determined using a three point quadrature of the thermodynamic integral (Simpson's rule). Further details on method and error estimation can be found in the Supporting Information.

In order to investigate the dependence of water adsorption energies on the number of TiO_2 layers, some static calculations were also performed. The same settings stated above were adopted, and ion configurations were optimized by using the BFGS minimizer. Water adsorption was applied to both surfaces of the slab symmetrically, resulting in the cancellation of surface dipoles.

3. Results and Discussion

3.1. Acidity of Surface Groups. The method of simultaneous deletion and insertion of protons was applied to the DFTMD model system depicted in Figure 1. We find an intrinsic pK_a of 7.8 for $TiOH_2$ and −1.9 for Ti_2OH^+. The statistical uncertainty in these estimates is approximately 2 pK_a units (see the Supporting Information) The pK_a's as entered in Table 1 have therefore been rounded off to pK_a = 8 for $TiOH_2$ and pK_a = −2 for Ti_2OH^+. Substitution in eq 6 yields a PZC of 3. If the thermochemical corrections for restraining potentials are applied (see the Supporting Information), these numbers increase by 1 pK_a unit to 9, −1, and 4, respectively.

Experimental data on the PZC of TiO_2 available from the literature are mostly based on measurements performed with rutile powder or polycrystalline samples. The results vary between 3 and 7 depending on sample preparation (synthesis) and electrolytes used. The consensus according to the compilation of ref 20 is a value of about 6. More recently, Bullard and Cima investigated the surface orientation

Table 1. Results for the Free Energies ΔA of Reactions 9 and 11 (referred to in the text as ΔA_1 respectively ΔA_2), the pK_a of $TiOH_2$ and Ti_2OH^+ Computed According to eqs 10 and 12, the Corresponding PZC (eq 6), ΔpK_a, and ΔA_{diss} [the dissociation free energy of adsorbed water] (numbers in parentheses are corrected for restraining potentials, see Supporting Information) for This Work (DFTMD), Using the MUSIC Model Taken from refs 3 and 13, and the SBE Model Taken from ref 4[a]

	DFTMD (this work)		MUSIC (refs 3, 13)		SBE (ref 4)	
	$TiOH_2$	Ti_2OH^+	$TiOH_2$	Ti_2OH^+	$TiOH_2$	Ti_2OH^+
ΔA (eV)	− 0.6	1.0				
pK_a	8 (9)	−2 (− 1)	7.5; 5.9	4.4; 4.9	8.2−9.3	2−3.7
PZC	3 (4)		6; 5.4		5.9−6.4	
ΔpK_a	10 (10)		3.1; 1		4.5−6.7	
ΔA_{diss} (eV)	0.6 (0.6)		0.18; 0.06		0.27−0.40	

[a] Only the PZC can be directly compared to the experiment. The relevant data here are the recent experimental estimates for the 110 surface of 5.5−4.8[23] and 4.8 ± 0.3[21] (see text).

dependence of the PZC of rutile TiO_2 using atomic force microscopy (AFM) and X-ray photoelectron spectroscopy (XPS) techniques.[23] For the rutile TiO_2(110) surface, they give a PZC of 5.5−4.8. This is consistent with the 4.8 ± 0.3 obtained by another study using second-harmonic generation spectroscopy (SHG).[21] We conclude therefore that our DFTMD result of 4 is in fair agreement with experiment, taking the computational error margin of 2 pK_a units (100 meV) into consideration.

As discussed in the Introduction, it is not possible to compare individual pK_a values directly to experiment. However a comparison with the predictions of models is of interest in its own right because of the uncertainties involved in setting up these models. The revised 1996 version of the MUSIC model[3] introduces two important refinements. The first is the use of actual bond valence computed from the length of the metal oxygen bonds. A second improvement over the older 1989 version[2] is the modeling of solvent effects which are taken into account by extending the expression for the actual bond valence with a term depending on hydrogen bonding. The relevant structural variables in this term are the number of hydrogen bonds donated (m) and accepted (n) by the base forms of the surface (hydr)oxide groups. m is fixed by the chemical species; $m = 0$ for Ti_2O and $m = 1$ for $TiOH^-$. To determine n the MUSIC model makes the assumption that $m + n = 2$, arguing that the total H coordination of a surface O ion is decreased by two compared to the coordination in liquid water.[3] One bond is replaced by the bond to the metal. Steric hindering eliminates a second bond. This means that $n = 2$ for Ti_2O and $n = 1$ for $TiOH^-$. With this hypothetical H-bond coordination and the actual bond valence determined by the experimental Ti−O distances of the bulk solid, the revised MUSIC model gives a pK_a of 7.5 for $TiOH_2$ and 4.4 for Ti_2OH^+ (these numbers are also listed in Table 1).[3] Compared to the DFTMD calculation, we see that the estimates for the pK_a of $TiOH_2$ are in good agreement. The fraction of dissociated terminal water molecules is predicted to be small under pH neutral conditions (pH = 7). A Ti_2O group is however significantly more basic (5 pK_a units) in the MUSIC model compared to the DFTMD calculation with a corresponding

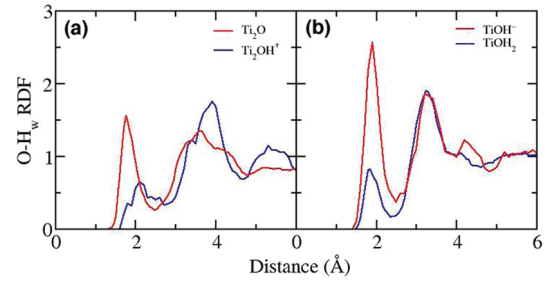


Figure 2. Radial distribution functions (RDF) between oxygen atoms of surface groups and hydrogen atoms in water. (a) The Ti_2OH^+/Ti_2O pair and (b) the $TiOH_2/TiOH^-$ pair. The deprotonated and protonated states are distinguished by red and blue. Coordination numbers obtained by integrating the first peaks are 1.1 for Ti_2O, 1.0 for Ti_2OH^+, 1.9 for $TiOH^-$, and 0.6 for $TiOH_2$, respectively.

shift of the PZC from 4 in the DFTMD calculation to 6 in the MUSIC model.

The critical structural parameters in the MUSIC model, the Ti−O bond lengths and number of hydrogen bonds, are directly accessible observables in molecular dynamics simulation. This suggests that the discrepancies between the pK_a values predicted by the MUSIC model and computed from the DFTMD simulation can be analyzed in more detail by comparing MD averages of Ti−O bond lengths and the number of hydrogen bonds to the values assumed in the MUSIC model. We can also substitute the MD estimates in the MUSIC model to see how the predicted values change. This was also the strategy followed by Machesky and co-workers in ref 13. Here, we will repeat some of this analysis using the results of our DFTMD simulation.

Coordination numbers are normally defined by the integral of the first peak of a radial distribution function (RDF). For an assessment of the variable n of the MUSIC model, the relevant RDFs are between the oxygen atoms in a Ti_2O and $TiOH^-$ surface group and hydrogen atoms in water. These RDFs are shown in Figure 2. The corresponding coordination numbers are $n = 1.1$ for Ti_2O and $n = 1.9$ for $TiOH^-$. MD results and values used in the revised MUSIC model are compared in Table 2. Consistent with the force field model of ref 13, DFTMD finds a higher coordination number for a terminal hydroxyl than assumed in the MUSIC model and a lower coordination for a bridging oxygen. The difference is 0.9 for both surface groups, corresponding to a change in pK_a of ∼3.6 units when substituted into the MUSIC model. More serious is that these changes go in the opposite direction, increasing the pK_a of Ti_2OH^+ from 4.4 to 7.9 and decreasing the pK_a of $TiOH_2$ from 7.5 to 4. The interchange in the order of the acidity, while having a minor effect on the PZC (eq 6), leads to a negative ΔpK_a of −4, implying that molecular absorption on TiO_2 is unstable (see eq 8).

The Ti−O bond lengths of surface groups are found to be on average somewhat shorter than the bulk values used in the MUSIC model. The precise values are 1.89 Å versus 1.95 Å for Ti_2O and 1.91 Å versus 1.98 for $TiOH_2$. The actual bond valence s is computed in the MUSIC model as $s = \exp(R - R_0)/b_0$, where R is the metal oxygen bond length, R_0 is a parameter specific to the solid oxide and $b = 0.37$ Å. The value for R_0 used in ref 3 for rutile is $R = 1.808$

Table 2. Estimation of Surface Acidity of TiO_2 Rutile (110) Using MUSIC and SBE Models from Structural Parameters Determined by the Present DFTMD Simulations[a]

	$TiOH_2/TiOH^-$			Ti_2OH^+/Ti_2O		
	m	n	pK_{a1}	m	n	pK_{a2}
this work	1	1.9	4(2)	0	1.1	8(6)
MUSIC model[3]	1	1	7.5	0	2	4.4

	r_{M-O}	s/r_{M-OH}	pK_{a1}	pK_{a2}	PZC	ΔpK_a
this work	1.91	0.2283	1.8	8.3	5.1	6.4
SBE model[4]	1.96	0.2248	2.0	8.4	5.2	6.4

[a] m and n as defined in the revised MUSIC model[3] are the number of donated and accepted hydrogen bonds by the conjugate base of surface (hydr)oxide groups. pK_{a1} and pK_{a2} are the recomputed acidities according to ref 3. The numbers in parentheses are the adjustments after the change to DFTMD bond length is taken into account (see text). For convenience, the original MUSIC predictions are carried over from Table 1. The structural input as determined by DFTMD for the SBE model[4] is given in the lower half of the table. r_{M-O} is the bond length in Å between the metal ion and O at the surface. r_{M-O} was determined in our simulation as the average Ti−O distance of a $TiOH^-$ group. s is the Pauling bond valence of the metal ion (2/3 for Ti^{4+} in TiO_2). The SBE model sets $r_{M-OH} = r_{M-O} + 1.01$. Parameterization of the SBE model depends on the choice of the double layer model. Here, we use the pK_a consistent with the triple layer model.[4]

Å. With these parameters, the reduction in bond length increases the s of a TiO bond by about 0.1 for both groups, corresponding to a decrease in pK_a of 2. Applying this adjustment to the acidities in Table 2, a terminal water becomes even more acidic ($pK_{a1} = 2$), widening the gap with a DFTMD value of 9 (Table 1). The adjusted pK_{a2} shows a similar discrepancy with the DFTMD estimate (see Table 2).

The conclusion must be that substitution of the observed DFTMD parameters leads to unrealistic intrinsic pK_a values which are not matching our estimates obtained from free energy calculation based on the same DFTMD simulation. While some major inaccuracy in the DFTMD approach cannot be completely excluded (such as finite size effects, see section also section 3.2), the observation that the use of the DFTMD hydrogen bond coordination numbers can lead to a qualitative change in the picture of the surface acidity of TiO_2 seems to support the criticism by Bickmore et al.[7,8] on the way solvation is treated in the MUSIC model (see also ref 6). In particular, these authors have questioned the direct coupling of the hydrogen bonding to the bond valence determining the undersaturation of basic oxygens as is characteristic of the MUSIC model.

Finally, we comment on the comparison to the SBE model also included in Tables 1 and 2. This comparison is of interest because the SBE approach to surface structure is more elementary, adopting the single site-two pK_a scheme.[4] Acidities correspond to the free energies of the protonation states of a generic site, namely, TiOH and $TiOH_2^+$ (see also section 2.1). pK_a in the SBE model is calculated from correlations with the electrostatic energy variable s/r_{M-OH}, where s is the formal Pauling bond valence charge of the metal ion and r_{M-OH} is the distance between metal ion and H atom in the OH group. The ionizable group resides at the interface between two dielectric continua, one representing the solvent and the other the solid. The model is directly fitted to experimental surface charge curves assuming certain double layer models.[4] Even though the identities of surface hydroxyl groups are ignored, pK_a's from the SBE model are fairly close to the numbers from our calculation and the MUSIC model. All three pK_a's of the basic component are very similar, while the SBE pK_a of the acidic component lies between the numbers from our DFTMD simulations and the MUSIC model (see Table 2).

Similar to the MUSIC model, we can also examine the response of the SBE model to exchange of the structural model parameters by the corresponding DFTMD averages. There is only one such parameter, namely, r_{M-OH}. In the model, this distance is evaluated from the equation $r_{M-OH} = r_{M-O} + 1.01$ where r_{M-O} is the length of the bond between the metal ion and O atom in the OH group. Similar to the MUSIC model, the SBE model takes the r_{M-O} value of the crystal. As mentioned, the time averaged value of r_{M-O} from our MD simulations is about 0.05 Å shorter (see also Table 2). Substituting this into the SBE model while keeping all other parameters constant yields almost the same pK_a's as the original SBE model. The success of the SBE model is remarkable considering its lack of structural and chemical detail. The microscopically inhomogeneous surface structure plays no role, and also the linear free energy relation is solely based on electrostatic interactions, ignoring other components of chemical bonding and replacing complicated relaxation effects at interfaces by a continuum medium model. Evidently, the combination of Pauling bond valence charge, electrostatics, and a double layer model in the SBE approach is capable of capturing all this complication in a simple linear relation, at least for the titania water interface studied here.

3.2. Dissociation Constant of Adsorbed Water. As discussed in section 2.1, the free energy change for dissociation of adsorbed water is related to the acidities of the two surface groups by eq 8. Substituting our calculated pK_a values of Table 1 into eq 8, we find $\Delta A_{diss} = 0.6$ eV. We estimate the statistical error in this result to be less than 0.2 eV (see the Supporting Information). Unlike the PZC, the thermochemical corrections applied to the two acidities end up canceling each other. The DFTMD result for ΔA_{diss} is compared in Table 1 to the corresponding free energies obtained from the MUSIC and SBE model acidities. All three dissociation free energies are positive, with the DFTMD estimate the largest. This is mainly because Ti_2OH^+ is more acidic according to DFTMD than predicted by the MUSIC and SBE models. A free energy cost for water dissociation of 0.6 eV strongly suggests that the reaction is unlikely to occur on perfect TiO_2(110). This appears to be in line with the recent work of Yamamoto et al.,[38] in which water adsorption was monitored by in situ XPS at ambient pressures and no noticeable water dissociation was observed except at O vacancies.

In assessing the DFTMD result for the dissociative adsorption of water, it is important to realize that ΔA_{diss} is known to be dependent on the number of TiO_2 layers in the model system. As has been verified repeatedly in the course of the numerous calculations under vacuum conditions,

Table 3. Variation of the Adsorption Energy (eV per molecule) of Water on Rutile TiO_2(110) in a Vacuum with the Number of TiO_2 Layers[a]

	0.5 ML		1 ML		
no. of TiO_2 layers	assoc.	diss.	assoc.	diss.	mix.
3	0.87	1.10	0.98	0.85	1.00
5	0.76	0.70	0.87	0.66	0.83
4 (from ref 35)[b]	0.86	0.64	0.82	0.63	0.77

[a] The water molecules complete the six fold coordination of the five fold coordinated surface Ti ions (terminal water molecules). Adsorption is symmetric on both surfaces of the slab under full geometry relaxation. (i) assoc. denotes associative adsorption, (ii) diss. means fully dissociative adsorption, and (iii) mix. is a mixed state with half water associatively adsorbed and half water dissociatively adsorbed (see refs 26 and 27 for a detailed description of water adsorption configurations). For comparison, the last row gives the adsorption energies of the corresponding 1 × 1 or 2 × 1 structures (whichever is the more stable) as obtained in ref 35 using the same density functional (PBE) applied to a somewhat different adsorption geometry (see further discussion in section 3.2). [b] H_2O adsorption on top layer only. To improve convergence with the number of layers, the bottom layer is terminated by fractional charges and has a constrained geometry.

adsorption energies on TiO_2 show a characteristic oscillation with the number of trilayers.[27,30,33,35] Also, the application of geometric constraints has a significant effect, which, when used appropriately, can accelerate the convergence.[33,35] Our results were calculated using five unconstrained trilayers of TiO_2 with double sided adsorption similar to the scheme employed in the gas-phase studies of Zhang and Lindan in refs 27 and 31. According to these authors, the accuracy of the adsorption energies computed with this approach is adequate for the estimation of relative stability of molecular and dissociative adsorption. In order to obtain a rough estimate of the bias introduced in our calculation by the limited thickness of the TiO_2 slab, we have carried out static test calculations of the adsorption energy of 0.5 and 1.0 monolayer (ML) of H_2O. The results are listed in Table 3 and compared to the corresponding PBE estimates of Kowalski and co-workers.[35] Their 4 trilayer adsorption energies are very close to the energies for a bulk surface as a result of the use of special termination and constraint methods. With the exception of one system (the 0.5 ML molecular adsorption), our five TiO_2 trilayer slab energies are higher by approximately 50 meV, consistent with the analysis in ref 35. We are therefore inclined to consider this 50 meV (corresponding to 1 pK_a unit) as a measure of our error due to the symmetric relaxed five layer geometry applied here with a better accuracy for relative absorption energies.

The results of Table 3 also confirm that 3 layers are not enough. This is best illustrated by the 0.5 ML system. This surface density is low enough to exclude hydrogen bonding between adsorbed water molecules complicating the adsorption energies. Dissociated water is considerably more stable than molecular water on three layers of TiO_2. For five layers of TiO_2, the stability is reversed. For the 1 ML surface coverage, 100% dissociation is not the energetically most favorable adsorption mode. Instead, a mixed state with associative and dissociative adsorption is preferred as a result of stabilization by intermolecular hydrogen bonding.[26] Note, however, that the data in Table 3 indicate that the tendency of a monolayer of water to dissociate on three layers of TiO_2 is still rather high.

These observations on the critical dependence of the stability of adsorbed water on the number of TiO_2 trilayers are in broad agreement with the extensive and detailed calculations on vacuum systems available from the literature.[27,30,33,35] The question is whether they can be carried over to models of TiO_2/H_2O interfaces as studied here. A quantitative investigation of the variation of surface pK_a with the slab thickness is forbiddingly expensive. However, a single MD run of a system of three layers of TiO_2 confirmed that this system retains its reactivity in bulk solution. We found that during the 10 ps trajectory up to about 20% of H_2O adsorbed on 5-fold Ti sites lost a proton to a nearby bridging oxygen. Water adsorbed on five layers of TiO_2 appears to be stable on this time scale. In this context, it is worth recalling that because of the finite temperature in a MD simulation (330 K, see section 2.3), a finite fraction of dissociated surface water molecules on three layers of TiO_2 does not necessarily mean that water dissociation is energetically more stable. This is true only if more than half of the surface waters dissociate in equilibrium.

A further issue requiring some comment is the comparison between the dissociation free energy determined from the estimate of surface pK_a in section 3.1 and the relative adsorption energies in Table 3. The $\Delta E_{ads} = 0.1-0.2$ eV difference in adsorption energy per molecule between molecular and dissociated monolayers is significantly smaller than the $\Delta A_{diss} = 0.6$ eV of Table 1. These two measures of the stability of water adsorbed on TiO_2 have however a rather different status. First of all, ΔE_{ads} is an enthalpy difference, while ΔA_{diss} is a free energy difference including entropy contributions. Furthermore, solvent effects in ΔE_{ads} can only arise due to hydrogen bonding in the first ad-layer. ΔA_{diss} also includes interactions with the second layer in the bulk solvent. However, the ≈0.5 eV difference between ΔA_{diss} and ΔE_{ads} is probably too large to be explained by these effects. More important is probably the difference in thermodynamic reference state. All of the water molecules in the calculation of ΔA_{diss} are molecular except those involved in the proton transfer. In contrast, in the calculation of ΔE_{ads}, half or all of the H_2O molecules are dissociated. ΔA_{diss} is therefore calculated under conditions approaching infinite dilution, while the solvent in the calculation of ΔE_{ads} is effectively a two-dimensional ionic solution at high ionic strength.

4. Conclusion

In summary, we have applied a recently developed DFTMD method for reversible proton insertion for a calculation of acidity constants of (hydr)oxide groups on the rutile Ti_2O(110) surface, i.e., bridge OH (Ti_2OH^+) and terminal H_2O adsorbed on 5-fold Ti sites ($TiOH_2$). Surface pK_a is estimated from the free energy of concerted protonation and deprotonation, equivalent to proton transfer between the surface and a H_3O^+ or H_2O in solution. The computed pK_a's of the two groups are −1 and 9, respectively, leading to a PZC of 4, which is within 2 pK_a units of the experimental

value for TiO_2 rutile (110). Using these two acidity constants, the free energy change of water dissociation at the TiO_2/H_2O interface has been determined as 0.6 eV. The positive free energy change indicates that water dissociation is not likely on a perfect Ti_2O(110) surface. While the discrepancy with the experiment for the PZC, the only observable directly accessible to experiment, is still (just) within the uncertainties in the calculation, an acidity of -1 for Ti_2OH^+ as obtained by DFTMD is likely an overestimation. This issue of the proton affinity of bridge oxygens clearly needs further investigation. In this context, a comparison to the 110 surface of SnO_2 which has the same structure as the TiO_2 surface studied here could be instructive.[34] This calculation is currently under way.

Analysis of the interfacial structure shows that some assumptions in the MUSIC model, in particular the number of hydrogen bonds to a bridging oxygen, are not justified. Using the DFTMD coordination numbers instead gave no improvement and in fact led to the prediction of a negative dissociation free energy of adsorbed water in contrast to the unambiguously positive dissociation free energy obtained in the DFTMD free energy calculation. These conflicting results can be seen as support for recent criticism of the way the MUSIC model couples explicit hydrogen bonding to the undersaturation determining the proton affinity.[7,8] This information may be useful for further development of models for intrinsic surface acidity constants.

Acknowledgment. We thank L.-M. Liu and A. Michaelides for useful discussions and showing their calculation data before publication. J.C. thanks M. Sulpizi for helpful discussions and M.-P. Gaigeot for help in computation of the vibrational spectrum from velocity autocorrelation functions. J. Vandevondele is acknowledged for technical support on CP2K. J.C. is grateful for EPSRC for financial support. The calculations for this work have been performed using an allocation of computer time on HECToR, the U.K.'s high-end computing resource funded by the Research Councils, as part of a grant to the UKCP consortium.

Supporting Information Available: A brief summary of our methodology for computation of free energies and some technical aspects are given, together with a description of thermochemical corrections. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Stumm, W. *Chemistry of the Soil-Water Interface*; Wiley: New York, 1992.

(2) Hiemstra, T.; Van Riemsdijk, W. H.; Bolt, G. H. *J. Colloid Interface Sci.* **1989**, *133*, 91.

(3) Hiemstra, T.; Venema, P.; Van Riemsdijk, W. H. *J. Colloid Interface Sci.* **1996**, *184*, 680.

(4) Sverjensky, D. A.; Sahai, N. *Geochim. Cosmochim. Acta* **1996**, *60*, 3773.

(5) Machesky, M. L.; Wesolowski, D. J.; Palmer, D. A.; Ridley, M. K. *J. Colloid Interface Sci.* **2001**, *239*, 314.

(6) Sahai, N. *Eviron. Sci. Technol.* **2002**, *36*, 445.

(7) Bickmore, B. R.; Tadanier, C. J.; Rosso, K. M.; Monn, W. D.; Egget, D. L. *Geochim. Cosmochim. Acta* **2004**, *68*, 2025.

(8) Bickmore, B. R.; Rosso, K. M.; Tadanier, C. J.; Bylaska, E. J.; Doud, D. *Geochim. Cosmochim. Acta* **2006**, *70*, 4057.

(9) Brown, I. D.; Altermatt, D. *Acta Crystallogr.* **1985**, *B41*, 244.

(10) Rustad, J. R.; Felmy, A. R.; Hay, B. J. *Geochim. Cosmochim. Acta* **1996**, *60*, 1563.

(11) Rustad, J. R.; Wasserman, E.; Felmy, A. R. *Surf. Sci.* **1999**, *424*, 28.

(12) Rustad, J. R.; Wasserman, E.; Felmy, A. R.; Wilke, C. *J. Colloid Interface Sci.* **1998**, *198*, 119.

(13) Machesky, M. L.; Predota, M.; Wesolowski, D. J.; Vlcek, L.; Cummings, P. T.; Rosenqvist, J.; Ridley, M. K.; Kubicki, J. D.; Bandura, A. V.; Kumar, N.; Sofo, J. O. *Langmuir* **2008**, *24*, 12331.

(14) Bandura, A. V.; Kubicki, J. D. *J. Phys. Chem. B* **2003**, *107*, 11072.

(15) Predota, M.; Bandura, A. V.; Cummings, P. T.; Kubicki, J. D.; Wesolowski, D. J.; Chialvo, A. A.; Machesky, M. L. *J. Phys. Chem. B* **2004**, *108*, 12049.

(16) Rustad, J. R.; Felmy, A. R.; Bylaska, E. J. *Geochim. Cosmochim. Acta* **2003**, *67*, 1001.

(17) Rustad, J. R.; Felmy, A. R. *Geochim. Cosmochim. Acta* **2005**, *69*, 1405.

(18) Sulpizi, M.; Sprik, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5238.

(19) Cheng, J.; Sulpizi, M.; Sprik, M. *J. Chem. Phys.* **2009**, *131*, 154504.

(20) Kosmulski, M. *J. Colloid Interface Sci.* **2002**, *253*, 77.

(21) Fitts, J. P.; Machesky, M. L.; Wesolowski, D. J.; Shang, X.; Kubicki, J. D.; Flynn, G. W.; Heinz, T. F.; Eisenthal, K. B. *Chem. Phys. Lett.* **2005**, *411*, 399.

(22) Zhang, Z.; Fenter, P.; Sturchio, N. C.; Bedzyk, M. J.; Machesky, M. L.; Wesolowski, D. J. *Surf. Sci.* **2007**, *601*, 1129.

(23) Bullard, J. W.; Cima, M. J. *Langmuir* **2006**, *22*, 10264.

(24) Bourikas, K.; Hiemstra, T.; Van Riemsdijk, W. H. *Langmuir* **2001**, *17*, 749.

(25) Bates, S. P.; Kresse, G.; Gillan, M. J. *Surf. Sci.* **1998**, *409*, 336.

(26) Lindan, P. J. D.; Harrison, N. H.; Gillan, M. J. *Phys. Rev. Lett.* **1998**, *80*, 762.

(27) Zhang, C.-J.; Lindan, P. J. D. *J. Chem. Phys.* **2003**, *118*, 4620.

(28) Zhang, C.-J.; Lindan, P. J. D. *J. Chem. Phys.* **2003**, *119*, 9183.

(29) Bandura, A. V.; Sykes, D. G.; Shapovalov, V.; Troung, T. N.; Kubicki, J. D.; Evarestov, R. A. *J. Phys. Chem. B* **2004**, *108*, 7844.

(30) Harris, L. A.; Quong, A. A. *Phys. Rev. Lett.* **2004**, *93*, 086105.

(31) Lindan, P. J. D.; Zhang, C.-J. *Phys. Rev. B* **2005**, *72*, 075439.

(32) Di Valentin, C.; Tilocca, A.; Selloni, A.; Beck, T. J.; Klust, A.; Batzill, M.; Losovyj, Y.; Diebold, U. *J. Am. Chem. Soc.* **2005**, *127*, 9895.

(33) Perron, H.; Vandenborre, J.; Domain, C.; Drot, R.; Roques, J.; Simoni, E.; Ehrhardt, J. J.; Catalette, H. *Surf. Sci.* **2007**, *601*, 518.

(34) Bandura, A. V.; Kubicki, J. D.; Sofo, J. O. *J. Phys. Chem. B* **2008**, *112*, 11616.

(35) Kowalski, P. M.; Meyer, B.; Marx, D. *Phys. Rev. B* **2009**, *79*, 115410.

(36) Henderson, M. A. *Surf. Sci.* **1996**, *355*, 151.

(37) Diebold, U. *Surf. Sci. Rep* **2003**, *48*, 53.

(38) Yamamoto, S.; Bluhm, H.; Andersson, K.; Ketteler, G.; Ogasawara, H.; Salmeron, M.; Nilsson, A. *J. Phys.: Condens. Matter* **2008**, *20*, 184025.

(39) Blumberger, J.; Tavernelli, I.; Klein, M. L.; Sprik, M. *J. Chem. Phys.* **2006**, *124*, 064507.

(40) Tateyama, Y.; Blumberger, J.; Sprik, M.; Tavernelli, I. *J. Chem. Phys.* **2005**, *122*, 234505.

(41) Blumberger, J. *J. Am. Chem. Soc.* **2008**, *130*, 16065.

(42) Adriaanse, C.; Sulpizi, M.; VandeVondele, J.; Sprik, M. *J. Am. Chem. Soc.* **2009**, *131*, 6046.

(43) VandeVondele, J.; Sulpizi, M.; Sprik, M. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 1936.

(44) Costanzo, F.; Sulpizi, M.; Della Valle, R. G.; Sprik, M. *J. Chem. Theory Comput.* **2008**, *4*, 1049.

(45) Yang, W.; Bitetti-Putzer, R.; Karplus, M. *J. Chem. Phys.* **2004**, *120*, 9450.

(46) Kleinman, L. *Phys. Rev. B* **1981**, *24*, 7412.

(47) Hunt, P.; Sprik, M. *Comput. Phys. Commun.* **2005**, *6*, 1805.

(48) Ayala, R.; Sprik, M. *J. Phys. Chem. B* **2008**, *112*, 257.

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

(50) The CP2K developers group. http://cp2k.berlios.de (accessed Feb 2010).

(51) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103.

(52) Lippert, G.; Hutter, J.; Parrinello, M. *Mol. Phys.* **1997**, *92*, 477.

(53) VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2003**, *118*, 4365.

(54) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703.

(55) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641.

(56) VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2007**, *127*, 114105.

(57) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **2005**, *122*, 014515.

CT100013Q

JCTC Journal of Chemical Theory and Computation

Implementation of Molecular Dynamics and Its Extensions with the Coarse-Grained UNRES Force Field on Massively Parallel Systems: Toward Millisecond-Scale Simulations of Protein Structure, Dynamics, and Thermodynamics

Adam Liwo,*[,†,‡] Stanisław Ołdziej,[‡,§] Cezary Czaplewski,[†,‡] Dana S. Kleinerman,[‡] Philip Blood,[||] and Harold A. Scheraga[‡]

Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland, Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, Laboratory of Biopolymer Structure, Intercollegiate Faculty of Biotechology, University of Gdańsk, Medical University of Gdańsk, Kładki 24, 80-822 Gdańsk, Poland, and Pittsburgh Supercomputing Center, Carnegie Mellon University, University of Pittsburgh, 300 S. Craig Street, Pittsburgh, Pennsylvania 15213

Received August 4, 2009

Abstract: We report the implementation of our united-residue UNRES force field for simulations of protein structure and dynamics with massively parallel architectures. In addition to coarse-grained parallelism already implemented in our previous work, in which each conformation was treated by a different task, we introduce a fine-grained level in which energy and gradient evaluation are split between several tasks. The Message Passing Interface (MPI) libraries have been utilized to construct the parallel code. The parallel performance of the code has been tested on a professional Beowulf cluster (Xeon Quad Core), a Cray XT3 supercomputer, and two IBM BlueGene/P supercomputers with canonical and replica-exchange molecular dynamics. With IBM BlueGene/P, about 50% efficiency and a 120-fold speed-up of the fine-grained part was achieved for a single trajectory of a 767-residue protein with use of 256 processors/trajectory. Because of averaging over the fast degrees of freedom, UNRES provides an effective 1000-fold speed-up compared to the experimental time scale and, therefore, enables us to effectively carry out millisecond-scale simulations of proteins with 500 and more amino acid residues in days of wall-clock time.

1. Introduction

Simulations of conformational changes in proteins and dynamics of protein conformations are nowadays of great importance in biochemistry, biophysics, and medical sciences.[1−13] All-atom molecular dynamics (MD) *ab initio* folding simulations (which require at least a microsecond time scale) are still restricted to the nanosecond time scale for large proteins and are possible only for proteins with lengths up to 60 residues with implicit-solvent approaches (however, simulations of the dynamics of big proteins starting from the experimental structure have been carried out since the early 1990s[6]), although great progress has been made with distributed computing (the FOLDING@HOME project);[14] the creation of very efficient load-balanced parallel codes such as GROMACS,[15] NAMD,[16] or DESMOND;[17] and, very recently, with the implementation of all-atom MD programs

* Corresponding author phone: +48 58 523 5430, fax: +48 58 523 5472, e-mail: adam@chem.univ.gda.pl.
† Faculty of Chemistry, University of Gdańsk.
‡ Cornell University.
§ Medical University of Gdańsk.
|| Pittsburgh Supercomputing Center.

10.1021/ct9004068
Published on Web 02/08/2010

on graphical processor units (GPUs)[18] and the construction of dedicated machines.[19] For recent developments of the software for all-atom MD simulations, see the latest review by Klepeis et al.[20] Coarse-grained models have, therefore, become of great importance in the field.[21−26] In the past decade, we have been developing a physics-based model of polypeptide chains, hereafter referred to as UNRES (for UNited RESidue).[27−39]

UNRES has been applied with considerable success in energy-based protein-structure prediction[40] and was later extended[41−43] to simulations of protein-folding pathways by implementing a coarse-grained dynamics approach, providing a 4000-fold speed-up compared to all-atom MD with an explicit solvent.[42,43] This speed-up factor was obtained by comparing the mean first passage time (MFPT) of the simulated folding of deca-alanine obtained with UNRES to that obtained with AMBER with explicit water. However, this speedup is not as big for larger proteins for which the water layer constitutes a smaller fraction of the system; also, introducing a cutoff on nonbonded interactions in all-atom simulations (which has not yet been performed for UNRES) causes a major reduction of the computation time. With the UNRES/MD approach, *ab initio* folding simulations of 75-residue proteins are possible within hours of single-processor time[43] (compared to weeks for all-atom MD with implicit solvent). Effectively, proteins fold in nanoseconds[42,43] with UNRES, while the shortest experimental protein-folding times are on the order of microseconds.[44] Consequently, UNRES provides a ∼1000-fold speed-up with respect to the experimental time scale. It should be noted that the coarse graining of protein representation, in particular using implicit solvent, reduces the accuracy of simulations; however, they still enable us to draw meaningful conclusions regarding protein folding. Trivial parallelization of molecular dynamics simulations, in which a single trajectory is calculated independently by a given processor, has enabled us[45] to study the folding kinetics of protein A, without biasing the potential toward the native structure, as in similar other studies.[46,47] With the aid of replica-exchange (REMD)[48] and multiplexed replica exchange molecular dynamics (MREMD),[49] we are now able to simulate protein-folding thermodynamics and perform ensemble-based prediction of protein structure.[36,37] We parallelized these algorithms,[50,51] and they scale well even for thousands of trajectories. Because of infrequent communication characteristic of the REMD and MREMD algorithms, they perform equally well on Beowulf clusters and on supercomputers with a fast interconnect. However, even with this reduction of computational time, *ab initio* folding simulations of proteins with a size of about 150 amino acid residues take weeks, and those of larger proteins are still out of reach in reasonable time, if a single processor handles the energy and gradient evaluation of a given conformation.

In this paper, we report the extension of UNRES to large protein systems, by introducing fine-grained parallelization of energy and gradient evaluation and other elements of an elementary step of molecular dynamics, in addition to *coarse-grained* parallelization, which controls it; this means that a group of fine-grained tasks is included in a coarse-grained task, which is dedicated to a given conformation. We demonstrate that, by setting 50% efficiency as a reasonable cutoff for the performance of a fine-grained-parallelized algorithm, it is reasonable to run calculations involving fine-grained parallelization with 8−16 processors dedicated to a single conformation for proteins with a size of more than 60 amino acid residues on professional Beowulf clusters and on the Cray XT3, which have faster processors (and thus less computation time per processor), and, for proteins with 30 amino acid residues or more, calculations on the IBM BlueGene/P, which has slower processors (and thus more computation time per processor) together with a fast interconnect, are reasonably efficient. For proteins with a size of about 800 amino acid residues, the achievable speedup is about 20-fold with professional Beowulf clusters, 30-fold with the Cray XT3, and over 100-fold with IBM BlueGene/P. This enables us to run *ab initio* simulations of folding dynamics and thermodynamics of such proteins in days of wall-clock time provided that massively parallel resources are available.

This paper is organized as follows. In sections 2.1, 2.2, and 2.3, we give a brief summary of the UNRES force field and MD and its extensions with UNRES. In section 2.4, we characterize the programming environment and machines used in this study. In section 2.5, we describe the parallelization scheme and its implementation. In section 3, we report the performance of the fine-grained code, and finally, in section 4, we discuss the implications of our work in biomolecular simulations and possible extensions of the approach, including implementation of the code on the GPUs.

2. Methods

2.1. The UNRES Force Field. In the UNRES model,[27−39] a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p). Each united peptide group is located in the middle between two consecutive α-carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α-carbons serving only to define the chain geometry, as shown in Figure 1. The UNRES force field has been derived as a restricted free energy (RFE) function[29,30] of an all-atom polypeptide chain plus the surrounding solvent, where the all-atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (viz., the degrees of freedom of the solvent, the dihedral angles χ for rotation about the bonds in the side chains, and the torsional angles λ for rotation of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual bonds).[52] The RFE is further decomposed into factors coming from interactions within and between a given number of united interaction sites.[30] Expansion of the factors into generalized Kubo cumulants[53] enabled us to derive approximate analytical expressions for the

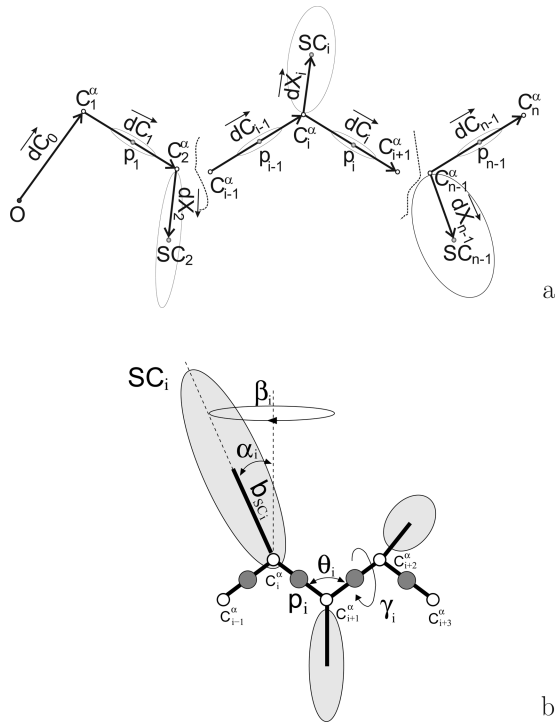


Figure 1. The UNRES model of the polypeptide chains. (a) Illustration of the variables used in MD. The terminal residues and a residue inside the chain are shown. The C^α atoms, the centers of the peptide groups (p), and side-chain centers (SC) are indicated by small open circles, while open ellipses centered at the p and SC centers indicate that these sites interact through noncentral forces. The C^α atoms serve only as geometric points and are not interaction sites. The peptide groups are positioned halfway between two consecutive C^α's. The geometry of the chain is defined by the vector $\mathbf{d}\vec{\mathbf{C}}_0$, defining the position of the first C^α atom (pointing from the origin O of the reference system to this atom), by backbone virtual bonds ($\mathbf{d}\vec{\mathbf{C}}_1$, $\mathbf{d}\vec{\mathbf{C}}_2$, ..., $\mathbf{d}\vec{\mathbf{C}}_{n-1}$), and side-chain ($\mathbf{d}\vec{\mathbf{X}}_2$, $\mathbf{d}\vec{\mathbf{X}}_3$, ..., $\mathbf{d}\vec{\mathbf{X}}_{n-1}$) vectors. For Gly residues, the respective SC atoms are methylene groups and are located exactly at the corresponding C^α atoms; therefore, there are no $\mathbf{d}\vec{\mathbf{X}}$ vectors for these residues. The C_1^α or C_n^α atoms are either the terminal methyl groups (in these cases, they are treated as Gly "side chains") if a chain is terminally blocked or dummy atoms if no blocking groups are present. (b) Illustration of internal coordinates pertaining to the ith residue used in eq 1: backbone virtual-bond-valence angles (θ_i), backbone virtual-bond-dihedral angle (γ_i), side-chain virtual-bond length (b_{SC_i}), and the angles α_{SC_i} and β_{SC_i} defining the position of the ith side chain with respect to the local coordinate frame defined by C_{i-1}^α, C_i^α, and C_{i+1}^α. All peptide groups are assumed to be in a trans configuration with an equilibrium virtual-bond length of 3.8 Å.

respective terms,[29,30] including the *multibody* or *correlation* terms, which are derived in other force fields from structural databases or on a heuristic basis.[54] The theoretical basis of the force field is described in detail in our earlier paper.[30]

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

$$
\begin{aligned}
U = {} & w_{SC}\sum_{i<j} U_{SC_iSC_j} + w_{SCp}\sum_{i\neq j} U_{SC_ip_j} + \\
& w_{pp}^{VDW}\sum_{i<j-1} U_{p_ip_j}^{VDW} + w_{pp}^{el} f_2(T)\sum_{i<j-1} U_{p_ip_j}^{el} + \\
& w_{tor} f_2(T)\sum_i U_{tor}(\gamma_i) + w_{tord} f_3(T)\sum_i U_{tord}(\gamma_i,\gamma_{i+1}) + \\
& w_b\sum_i U_b(\theta_i) + w_{rot}\sum_i U_{rot}(\alpha_{SC_i},\beta_{SC_i}) + \\
& w_{bond}\sum_i U_{bond}(d_i) + w_{corr}^{(3)} f_3(T) U_{corr}^{(3)} + w_{corr}^{(4)} f_4(T) U_{corr}^{(4)} + \\
& w_{corr}^{(5)} f_5(T) U_{corr}^{(5)} + w_{corr}^{(6)} f_6(T) U_{corr}^{(6)} + w_{turn}^{(3)} f_3(T) U_{turn}^{(3)} + \\
& w_{turn}^{(4)} f_4(T) U_{turn}^{(4)} + w_{turn}^{(6)} f_6(T) U_{turn}^{(6)} \qquad (1)
\end{aligned}
$$

where θ_i is the backbone virtual-bond angle, γ_i is the backbone virtual-bond-dihedral angle, α_i and β_i are the angles defining the location of the united side-chain center of residue i (Figure 1b), and d_i is the length of the ith virtual bond, which is either a $C^\alpha\cdots C^\alpha$ virtual bond or a $C^\alpha\cdots SC$ virtual bond. Each term is multiplied by an appropriate weight, w_x, and the terms corresponding to factors of an order higher than 1 are additionally multiplied by the respective temperature factors, which were introduced in our recent work[36] and which reflect the dependence of the first generalized-cumulant term in those factors on temperature, as discussed in refs 36 and 55. The factors f_n are defined by eq 2.

$$
f_n(T) = \frac{\ln[\exp(1) + \exp(-1)]}{\ln\{\exp[(T/T_o)^{n-1}] + \exp[-(T/T_o)^{n-1}]\}} \qquad (2)
$$

where $T_o = 300$ K.

The term $U_{SC_iSC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains, which implicitly contain the contributions from the interactions of the side chain with the solvent. The term $U_{SC_ip_j}$ denotes the excluded-volume potential of the side-chain−peptide-group interactions. The peptide-group interaction potential is split into two parts: the Lennard-Jones interaction energy between peptide-group centers ($U_{p_ip_j}^{VDW}$) and the average electrostatic energy between peptide-group dipoles ($U_{p_ip_j}^{el}$); the second of these terms accounts for the tendency to form backbone hydrogen bonds between peptide groups p_i and p_j. The terms U_{tor}, U_{tord}, U_b, U_{rot}, and U_{bond} are the virtual-bond-dihedral angle torsional terms, virtual-bond dihedral angle double-torsional terms, virtual-bond angle bending terms, side-chain rotamer, and virtual-bond-deformation terms; these terms account for the local propensities of the polypeptide chain. The terms $U_{corr}^{(m)}$ represent correlation or multibody contributions from the coupling between backbone-local and backbone-electrostatic interactions, and the terms $U_{turn}^{(m)}$ are correlation contributions involving m consecutive peptide groups; they are, therefore, termed turn contributions. In the present version of UNRES, the terms of an order higher than 4 are not present because, in our earlier work,[34] we found that the correlation terms of an order up to 4 are sufficient to reproduce regular secondary structures, while including higher-order terms results in a substantial increase of the cost of energy and force evalua-

tion. The multibody terms are indispensable for reproduction of regular α-helical and β-sheet structures.[29,30,54]

The internal parameters of $U_{p_ip_j}^{VDW}$, $U_{p_ip_j}^{el}$, U_{tor}, U_{tord}, U_b, $U_{corr}^{(m)}$, and $U_{turn}^{(m)}$ were derived by fitting the analytical expressions to the RFE surfaces of model systems computed at the MP2/6-31G** *ab initio* level;[32,33,35] the parameters of U_{bond} and U_{rot} were derived similarly from the energy surfaces of terminally blocked amino acid residues calculated with the AM1 semiempirical method,[38,39] while the parameters of $U_{SC_iSC_j}$ and $U_{SC_ip_j}$ were derived by fitting the calculated distribution functions to those determined from the PDB.[28] The *w*'s are the weights of the energy terms, and they can be determined only by optimization of the potential-energy function, as described in our earlier work.[31,34,36] In the present study, we use a recent version of UNRES developed by extensive random search of the energy-term weights.[56]

2.2. Molecular Dynamics with UNRES. In this section, we summarize briefly the theory behind the UNRES/MD approach and recall the equations necessary to understand its algorithmic aspects. The reader is referred to our earlier work[41,42] for details. The equations of motion for the UNRES chain are Langevin-dynamics equations because the solvent is implicit in UNRES. Consequently, it contributes to conservative forces (through the RFE) and gives rise to nonconservative forces which originate in energy exchange of the polypeptide chain with the solvent (the stochastic and friction forces). Because the geometry of an UNRES chain is not uniquely defined by the Cartesian coordinates of the interacting sites, we chose the virtual-bond vectors $C_i^\alpha \cdots C_{i+1}^\alpha$ (denoted as $\mathbf{dC}_i$) and $C_i^\alpha \cdots SC_i$ (denoted as $\mathbf{dX}_i$) as generalized coordinates, as expressed by eq 3.

$$\mathbf{q} = (\mathbf{dC_0}, \mathbf{dC_1}, ..., \mathbf{dC}_{n-1}, \mathbf{dX}_s, \mathbf{dX}_{s+1}, ..., \mathbf{dX}_e)^T \quad (3)$$

where *s* and *e* are the indices of the first and of the last side-chain vectors, respectively (i.e., of the first and of the last nonglycine residues; the glycine residues carry no **dX** vectors). The vector $\mathbf{dC_0}$ specifies the position of the first C^α atom of the chain. The vectors $\dot{\mathbf{q}}$ and $\ddot{\mathbf{q}}$ denote generalized velocities and generalized accelerations, respectively We assume that the virtual bonds are elastic rods with mass distribution that scales with the length of a rod.[41] The Cartesian coordinates of the interacting sites, p_i and SC_i, defined by eq 4

$$\mathbf{x} = (\mathbf{r}_{p_1}, \mathbf{r}_{p_1}, ..., \mathbf{r}_{p_{n-1}}, \mathbf{r}_{SC_1}, \mathbf{r}_{SC_2}, ..., \mathbf{r}_{SC_n})^T \quad (4)$$

are related to the generalized coordinates by a linear transformation expressed by eq 5

$$\mathbf{x} = \mathbf{Aq} \quad (5)$$

where **A** is a constant matrix such that $a_{i(k),j} = 0$ [$i(k)$ being a Cartesian coordinate of site *k*] if the coordinates up to *j* correspond to virtual-bond vectors of the part of the chain to the right of site *k*, $a_{i(k),j} = 1$ if the coordinate corresponds to virtual-bond vector to the left of site *k* or to a $C^\alpha \cdots SC$ virtual bond containing the side chain with index *k*, and $a_{i(k),j} = 1/2$ if the coordinate corresponds to the virtual-bond vector containing the peptide group with index *i*, as expressed by eq 6.[41] The same relationship holds between the time derivatives of **x** and **q**.

$$\mathbf{A} = \begin{pmatrix} \mathbf{1} & \frac{1}{2}\mathbf{1} & & & & & & & & \\ \mathbf{1} & \mathbf{1} & \frac{1}{2}\mathbf{1} & & & & & & & \\ \vdots & \vdots & \vdots & \ddots & & & & & & \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \dots & \frac{1}{2}\mathbf{1} & & & & & \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & h_{SC_s}\mathbf{1} & & & \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & h_{SC_{s+1}}\mathbf{1} & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \dots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \dots & h_{SC_e}\mathbf{1} \end{pmatrix} \quad (6)$$

where h_{SC_i} is 0 if SC_i = Gly and is 1 otherwise, **1** is a 3 × 3 identity matrix, and **0** is a 3 × 3 matrix composed of zeros. It should be noted that the columns of matrix **A** run through all variables [the elements of the vector **q** of eq 3], and the rows run through the Cartesian coordinates of the interacting sites.

The peptide groups and side chains are represented as stretchable rods with uniformly distributed masses.[41] The Langevin equation for UNRES[42] is given by eq 7

$$\mathbf{F} = \mathbf{G}\ddot{\mathbf{q}} = -\nabla_{\mathbf{q}}U(\mathbf{q}) - \mathbf{A}^T\Gamma\mathbf{A}\dot{\mathbf{q}} + \mathbf{A}^T\mathbf{f}^{rand} \quad (7)$$

$$\mathbf{G} = \mathbf{A}^T\mathbf{MA} + \mathbf{H} \quad (8)$$

where **M** is the diagonal matrix of the masses of the sites (united peptide groups and united side chains) defined by eq 9, **H** [a diagonal matrix defined by eq 10] is the part of the inertia matrix corresponding to the internal stretching motion of the virtual bonds,[41] **G** [defined by eq 8] is the inertia matrix, Γ is the diagonal friction tensor (represented by the friction matrix with elements γ_{ii} determined by Stokes law) acting on the interacting sites such that γ_{ii} is the coefficient of the site corresponding to the *i*th coordinate, $\mathbf{f}^{rand}$ is the vector of random forces acting on interacting sites, *U* is the UNRES potential energy defined by eq 1, $\nabla_{\mathbf{q}}$ denotes the gradient with respect to **q**, and **F** is the vector of forces. In this work, we did not use the stochastic and friction terms; therefore constant temperature was maintained by using the Berendsen thermostat[57] implemented in UNRES/MD in our earlier work.[42]

$$\mathbf{M} = \begin{pmatrix} m_p\mathbf{1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & m_p\mathbf{1} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & m_{SC_s}\mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & m_{SC_{s+1}}\mathbf{1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & m_{SC_e}\mathbf{1} \end{pmatrix} \quad (9)$$

where m_p is the mass of a peptide group and m_{SC_i} is the mass of the *i*th side chain.

The matrix **H** is defined by eq 10.

$$\mathbf{H} = \begin{pmatrix} I_p\mathbf{1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & I_p\mathbf{1} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & I_{SC_{i_1}}\mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & I_{SC_{i_2}}\mathbf{1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & I_{SC_{i_m}}\mathbf{1} \end{pmatrix} \tag{10}$$

with[42]

$$I_p = (1/12)m_p \text{ and } I_{SC_i} = (1/3)m_{SC_i} \tag{11}$$

To compute the accelerations from forces by using eq 7 requires matrix multiplication of the right-hand side of this equation by $\mathbf{G}^{-1}$ [eq 12].

$$\ddot{\mathbf{q}} = \mathbf{G}^{-1}\mathbf{F} \tag{12}$$

where **F** is the vector of forces defined by the right-hand side of eq 7. Because **G** is a constant matrix, its inverse [which involves $\mathcal{O}(n^3)$ computational effort] can be computed only once at the beginning of a simulation and used later. Moreover, it can be seen from eqs 6, 9, and 10 that, if the coordinates in vectors **x** and **q** are rearranged to group all x, then all y, and finally, all z coordinates and the rows and columns of the matrix **G** are rearranged accordingly, matrix **G** is separated into three blocks corresponding to x, y, and z parts, thereby reducing the number of operations in matrix multiplication by a factor of 3.

2.3. Replica-Exchange and Multiplexing-Replica-Exchange Molecular Dynamics. To sample the conformational space more efficiently than by canonical MD, we extended[50,51] the UNRES/MD approach with the replica-exchange[48] and multiplexing replica-exchange[49] molecular dynamics method. In this method, M canonical MD simulations are carried out simultaneously, each one at a different temperature. Initially the temperatures increase with the sequential number of simulations (trajectories). After every $m < M$ step, an exchange of temperatures between neighboring trajectories (in the order from 1 to M) is attempted, the decision about the exchange being made based on the Metropolis criterion which, taking into account the temperature dependence of the force field, is expressed by eq 13.

$$\Delta = [\beta_{i+1}U(\mathbf{X}_{i+1},\beta_{i+1}) - \beta_i U(\mathbf{X}_{i+1},\beta_i)] - [\beta_{i+1}U(\mathbf{X}_i,\beta_{i+1}) - \beta_i U(\mathbf{X}_i,\beta_i)],\ i = 1, 2, ..., M \tag{13}$$

where $\beta_i = 1/RT_i$, T_i being the absolute temperature corresponding to the ith trajectory, and $\mathbf{X}_i$ denotes the variables of the UNRES conformation of the ith trajectory at the attempted exchange point. If $\Delta \leq 0$, T_i and T_{i+1} are exchanged; otherwise the exchange is performed with probability $\exp(-\Delta)$.

The multiplexing variant of the REMD method (MREMD)[49] differs from the REMD method in that several trajectories are run at a given temperature. Each set of trajectories run at a different temperature constitutes a *layer*. Exchanges are attempted not only within a single layer but also between layers. In our very recent study,[51] we demonstrated that such a procedure increases the power of REMD very considerably, and convergence of the thermodynamic quantities is achieved much faster.

2.4. Programming Languages, Libraries, Other Software Used, and Machines. As in our previous work on the implementations of the conformational space annealing[58,59] and the REMD[50] and MREMD[51] versions of UNRES, the Message Passing Interface (MPI) library was implemented to construct the parallel code. Therefore, the description of parallelization in section 2.5 is message-passing oriented. The code was written in our laboratory, using the FORTRAN 77 programming language, and was parallelized using the MPI library. Standard blocking MPI_SCATTER and MPI_GATHER MPI library routines were implemented to send the data from a master task to the slave task and to collect data from the slave tasks, respectively, and MPI_REDUCE with the MPI_SUM operator was used to sum the contributions to energy, energy gradients, and results of matrix-vector multiplications from slave tasks at a fine-grain master task. The coarse-grained tasks and each of the fine-grained tasks formed separate MPI communicators to avoid possible message interception.[60]

For storage of trajectory files, we use the freely available Europort Data Compression XDRF library (http://hpcv100.rc.rug.nl/xdrfman.html). The BLAS routines[61] are used for matrix diagonalization.

The UNRES code with MD and MREMD has been implemented and tested on the Xeon Quad Core professional Beowulf cluster at the Academic Computer Center in Gdańsk, TASK (galera.task.gda.pl; 1344 Intel Xeon Quad-Core processors, 5376 cores, Mellanox InfiniBand interconnect with 20 Gb/s bandwidth; Web page at http://www.task.gda.pl/english/kdm.html), on the Cray XT3 supercomputer at the Pittsburgh Supercomputer Center [bigben.psc.edu; 4096 processors (dual-core), the 3D torus network; Web page at http://www.psc.edu/general/hardware.php], and on the following two IBM BlueGene/P massively parallel machines: at Argonne National Laboratory (intrepid.alcf.anl.gov; 40 960 quad-core compute nodes; Web page at http://www.alcf.anl.gov/resources/storage.php) and at the Jülich Supercomputer Center (jugene.fz-juelich.de; 73 728 compute nodes, each with a four-way 32-bit PowerPC 450 core 850 MHz processor, three-dimensional torus network; Web page at http://www.fz-juelich.de/jsc/jugene). Of these four machines, galera.task.gda.pl has the largest single-processor speed, while both IBM BlueGene/P machines have the fastest and the most efficient communication but the slowest processors.

2.5. Parallelization and Parallel Programming System Used. *2.5.1. General Parallelization Scheme.* As mentioned in the Introduction, the code is parallelized at a two-grain level:

(1) Coarse-grained level, in which a given coarse-grained task (CG) is assigned to a given trajectory. In a multitrajectory canonical molecular dynamics run, there is no communication between tasks except synchronization at the end

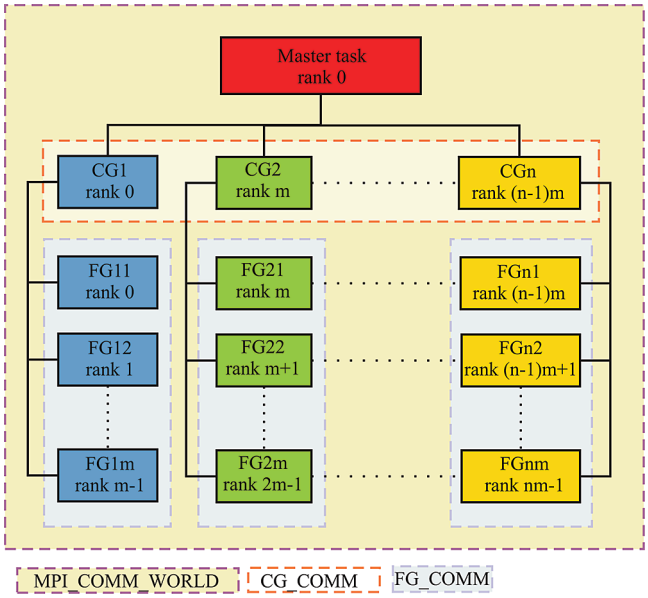


Figure 2. Scheme for the distribution of *n* coarse-grained (CG1, CG2, ..., CGn) and *m* fine-grained tasks for each coarse-grained task (FG11, FG12, ...FG1m, FG 21..., FGnm) among *nm* processes (with ranks 0, ..., nm − 1) and of communicators (shown as boxes bordered by dashed lines) pertaining to the respective tasks. MPI_COMM_WORLD is the communicator containing all processes. CG_COMM contains the processes assigned to coarse-grained tasks, and FG_COMM is the communicator containing processes assigned to a given fine-grained task.

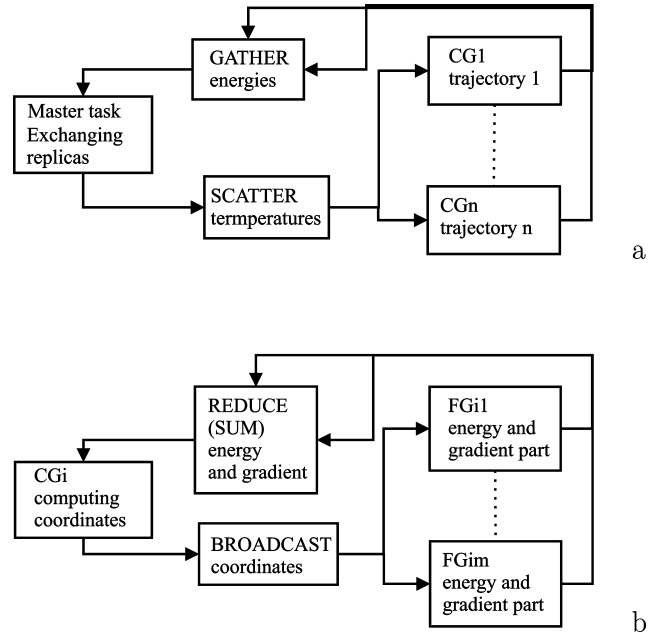


Figure 3. Basic operations and communication between the master task and the CG tasks (a), and between a CG task and the pertinent FG tasks (b).

of the job. For REMD, communication occurs at the time of exchanging replicas (see section 2.3).

(2) Fine-grained level, in which (i) computation of energy and forces, and (ii) computation of accelerations from forces [eq 7] are distributed to fine-grained tasks (FGs).

A scheme of tasks and processor assignment to tasks by rank is shown in Figure 2. Figure 3a and b shows the basic operations and communication between the master task and CG tasks in a replica-exchange run, and between a CG task and the pertinent FG tasks in energy and gradient evaluation.

The coarse-grained parallelism is the same as discussed in our earlier work;[50,51] therefore, we provide only a brief summary here.

As discussed in our earlier work,[51] one problem with the efficiency of parallel (M)REMD runs involving over 500 trajectories run simultaneously is the rapid loss of scalability because of synchronization implied in classical (M)REMD algorithms at each exchange. Even if the parallel system is homogeneous, 100% of processor time is never available for a compute task, which results in a slightly different processor speed, depending on the task. Moreover, a distance cutoff is applied in the computation of four-body correlation interactions,[29] and in our A-MTS algorithm[62] for integrating the equations of motion, the split number of a given time step depends on current conformation. Therefore, the computational effort depends on the trajectory. The discrepancy between the slowest and the fastest trajectories increases with increasing number of trajectories. Therefore, we modified[51] the (M)REMD algorithm to reduce synchronization overhead. As soon as a processor has executed the preset number of steps since the last exchange, it sends a message to the other processors to perform replica exchange. With this modification, we reached over 70% scalability for 4096 trajectories.[51]

The fine-grained parallelism introduced in this work is discussed in detail in the next subsections.

2.5.2. Complexity of Operations Involved in a Single MD Step. In order to decide about the parallelization strategy, the complexity of the operations involved in a single MD step must be analyzed. In all-atom MD simulations, almost all of the computational effort is connected with energy and force evaluation; however, in UNRES, the calculation of acceleration from forces involves a matrix−vector multiplication [eq 12], gradient transformation (discussed in section 2.5.4), and other operations specific for coarse-grained approaches.

The complexity and relative contributions to single-processor wall-clock time [calculated for the 1TF5 protein (767 residues, α + β)] of the energy terms and other operations performed during the production phase of an UNRES/MD run are summarized in Table 1. Because of the significant reduction of the number of interaction sites compared to an all-atom representation of polypeptide chains, we did not yet introduce a cutoff on nonbonded interactions, as opposed to all-atom MD software[15−17] (although we are planning to introduce the cutoff in the future). Therefore, the complexity of all long-range interactions is $\mathcal{O}(n^2)$, n being the number of residues (Table 1) and not $\mathcal{O}(n)$. It can be seen from Table 1 that the greatest effort is required to compute the average-electrostatic ($U^{el}_{p_ip_j}$) and van der Waals ($U^{VDW}_{p_ip_j}$) as well as third-order multibody ($U^{(3)}_{corr}$) interactions between the peptide groups; individual timings cannot be provided because these terms are computed in a single subroutine and use common intermediate quantities (such as, e.g., the distance between peptide-group centers). The next are the $U_{SC_iSC_j}$ and $U_{SC_ip_j}$ terms, and the least expensive

Table 1. Complexity and Timing of Different Operations of UNRES for the 1TF5 Protein Obtained with 1, 32, and 128 Processors at bigben.psc.edu[a]

operation	complexity	parallelized	% contribution		
			1 CPU	32 CPUs	128 CPUs
U_{bond}	$\mathcal{O}(n)$	yes	0.0	0.0	0.0
U_b	$\mathcal{O}(n)$	yes	0.4	0.2	0.1
U_{rot}	$\mathcal{O}(n)$	yes	0.2	0.1	0.1
U_{tor}	$\mathcal{O}(n)$	yes	0.1	0.0	0.0
U_{tord}	$\mathcal{O}(n)$	yes	0.5	0.0	0.1
$U_{turn}^{(3)}$	$\mathcal{O}(n)$	yes	0.0	0.0	0.0
$U_{turn}^{(4)}$	$\mathcal{O}(n)$	yes	0.1	0.1	0.1
$U_{SC_iSC_j}$	$\mathcal{O}(n^2)$	yes	21.2	9.5	5.4
$U_{SC_ip_j}$	$\mathcal{O}(n^2)$	yes	8.0	3.6	2.1
$U_{p_ip_j}^{el+VDW} + U_{corr}^{(3)}$	$\mathcal{O}(n^2)$	yes	62.6	29.3	18.6
$U_{corr}^{(4)}$[b]	$\mathcal{O}(n^2)$	yes	3.5	6.3	8.5
auxiliary to compute $U_{corr}^{(3)}$ and $U_{turn}^{(3,4)}$	$\mathcal{O}(n)$	attempted[d]	0.1	3.0	5.0
summing energy components[b]	$\mathcal{O}(n)$	yes	0.0	18.0	11.5
gradient transform[b]	$\mathcal{O}(n^2)$	yes	0.2	12.8	20.1
accelerations	$\mathcal{O}(n^2)$	yes	2.8	10.0	14.2
other	$\mathcal{O}(n)$	attempted[d]	0.3	7.1	14.2
total active[c] communication			N/A	16.7	28.2

[a] The operations were timed using TAU (http://www.cs.uoregon.edu/research/tau/home.php). [b] Includes the active communication and synchronization time due to load imbalance. [c] Does not include synchronization time due to load imbalance. [d] The corresponding code was parallelized, but communication overhead turned out to be greater in the present implementation than the gain from splitting the workload.

are the correlation terms; however, this is because we introduced a 7 Å distance cutoff on the computation of these interactions.[29] The single-body terms [of $\mathcal{O}(n)$ complexity] are by far less expensive. It can also be seen that, although energy and gradient computations constitute the dominant fraction of computation time with one processor, gradient transformation and computing accelerations from forces are operations of $\mathcal{O}(n^2)$ complexity and cannot be ignored in parallelizing the code. Moreover, the complexity of these operations cannot be reduced by introducing the cutoff.

2.5.3. Fine-Grained Parallelization of Energy- and Gradient-Component Calculations. The following three approaches are commonly implemented in the parallelization of energy and force calculations:[15,63,64] the particle- (atom)-decomposition approach also called the replicated-data approach, the domain- (spatial)-decomposition approach, and the force-decomposition approach.

In the particle-decomposition approach, N atoms are split between P processors, and each of the processors calculates the forces acting at the atoms assigned to it. The coordinates of all atoms must be known to all processors. The particle decomposition approach is naturally load-balanced if no cutoff is introduced on nonbonded interactions; when a cutoff is introduced, load-balancing can be achieved by randomizing the order of the atoms before dividing them between processors.[63,64] However, the memory requirements and cost of communication are on the order of $\mathcal{O}(N)$ (i.e., does not decrease with increasing number of processors). The all-to-all communication is required because a given processor must obtain the updated coordinates from all other processors. Moreover, because Newton's third law is usually implemented to reduce the amount of computations, the forces are exchanged between the processors, which requires additional communication.

In the domain-decomposition approach, the space occupied by the system is partitioned between processors, and each processor handles the particles which are contained in the region of space assigned to it. The load balance is easy to maintain only when the space is nearly uniformly filled with particles, as in all-atom simulations with explicit solvent carried out in a periodic solvent box; then each processor always handles approximately the same number of particles.[15,63,64] It is difficult to maintain the load balance for irregularly shaped systems, typical of simulations of proteins with implicit solvent, for which particle density varies significantly. The domain-decomposition algorithm implies the least memory requirement and communication cost; both are on the order of $\mathcal{O}(N/P)$. The all-to-all communication is avoided here because a given processor needs only the information from the processors that are assigned to the neighboring regions of space.

In the force-decomposition algorithm, the interatomic interactions are partitioned between processors in such a way that a given processor owns a block of force matrix pertaining to forces due to atoms from, say, i to j acting on atoms from, say, k to l.[63,64] Thus, this processor needs the coordinates of the atoms of only these two groups and the all-to-all communication is avoided. Consequently, the memory requirement and communication cost are on the order of $\mathcal{O}(N/\sqrt{P})$, i.e., between that of the particle- and domain-decomposition algorithms. As in the particle-decomposition algorithm, the load balance is natural with no cutoff on nonbonded interactions; with a cutoff, it can be maintained by randomizing the distribution of forces between the processors.

The UNRES energy terms of eq 1 and the corresponding forces can be divided into the following classes:

(1) The terms computed independently from the coordinates of UNRES objects within

(a) a single residue (U_{bond}; one-body terms)

(b) a number (up to 4) of neighboring residues (U_b, U_{rot}, U_{tor}, U_{tord}, $U_{turn}^{(3)}$, $U_{turn}^{(4)}$; local terms)

(c) pairs of residues ($U_{SC_iSC_j}$, $U_{SC_ip_j}$, $U_{p_ip_j}^{el}$, $U_{p_ip_j}^{VDW}$; pairwise terms)

(d) pairs of residues plus those of neighboring residue ($U_{corr}^{(3)}$; pairwise correlation terms)

(2) The terms which comprise two neighboring pairs of interacting peptide groups $p_i p_j$ and $p_{i+1} p_{j\pm1}$ ($U_{corr}^{(4)}$, $U_{corr}^{(5)}$, $U_{corr}^{(6)}$, $U_{turn}^{(6)}$).

As follows from Table 1, the bulk effort involves the computation of the pairwise terms included in 1c and 1d. Therefore, we designed energy and force parallelization to distribute these terms in the first place. In the present UNRES, no cutoff is applied on nonbonded interactions. Consequently, the workload corresponding to the computation of terms 1a−1d remains fixed, and each of these terms can be distributed to processors at the beginning of the calculations. It appears natural to use the particle- or force-decomposition approach. Given the fact that the solvent is implicit in UNRES and, therefore, particle density is not uniform, using the domain-decomposition approach would require a sophisticated algorithm to partition the space between processors (such as those designed for, e.g., solving the boundary-value problems with irregular shapes).[65] Moreover, the equations of motion are expressed in virtual-bond vectors in UNRES/MD, and therefore, communication with non-neighboring boxes could not be avoided because the forces expressed in UNRES interaction-site coordinates must be transformed to those expressed in virtual-bond vectors. Furthermore, the number of interaction sites is greatly reduced in UNRES compared to all-atom MD, and neither does storing all coordinates at each processor pose memory problems nor does sending the updated coordinates to all processors pose a significant communication problem.

Given the above considerations, we designed a parallelization scheme which is similar to force-decomposition algorithms in that the interactions and non-interaction sites are split between the processors, although we partition the upper triangular of the force matrix and not the whole force matrix. Thus, we make explicit use of Newton's third law. The decomposition of pairwise interactions is illustrated in Figure 4. Communication is not required to update the interaction list assigned to each processor. On the other hand, our scheme is similar to the particle-decomposition scheme in that the coordinates (virtual-bond vectors) are sent to all processors (Figure 2b). However, because only the FG master processor updates the virtual-bond vectors, this involves broadcast and not all-to-all operations.

The computation of the interactions included in group 1 was parallelized in the same manner; it should be noted, however, that the list of interactions consigned to a processor remains constant independent of the use of a cutoff. Although the complexity of the local interactions is only $\mathcal{O}(n)$, parallelizing this part of the code does not introduce any extra communication and, consequently, was implemented in UNRES.

Group 2 contains neighboring pairs of peptide groups, and quantities calculated in computing $U_{p_i p_j}^{el}$ and $U_{p_{i+1} p_{i\pm1}}^{el}$ are subsequently utilized in computing the correlation terms involving the respective pairs of peptide groups.[29,30] A pair of interactions necessary to compute a contribution to one of the UNRES energy terms listed in point 1d can happen to be split between two processors, and communication is, therefore, required to avoid multiple computation of the same quantities by different processors.

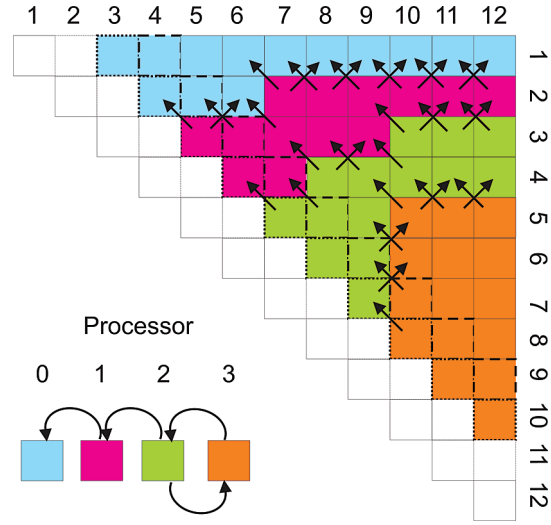


Figure 4. Illustration of domain decomposition of peptide-group interactions of a 13-residue chain (with 12 peptide groups) and exchange of interaction information to compute correlation interactions. A box in the upper-triangular array shown in the upper part of the graph corresponds to the interaction between peptide groups numbered by the row and the column of the array. Different colors mark assignment of the respective interactions to different processors, whose ranks are shown in the lower small panel. The white diagonal and next to diagonal boxes correspond to the same or neighboring peptide groups, for which long-range interactions are not considered. The dot-dashed and dashed lines lines border the 1,3 and 1,4 interactions, respectively. The arrows in the upper panel indicate the directions in which the interaction information is transmitted, and the half-circular arrows in the left lower panel show the flow of interaction information.

Because the quantities required to compute $U_{p_i p_{i+2}}^{el}$ and $U_{corr;\, p_i p_{i+2}}^{(3)}$ are also required for $U_{turn;\, i}^{(3)}$, and those required to compute $U_{p_i p_{i+3}}^{el}$ and $U_{corr;\, p_i p_{i+3}}^{(3)}$ are required to compute $U_{turn;\, i}^{(4)}$, computations of $U_{p_i p_{i+2}}^{el}$ and $U_{corr;\, p_i p_{i+2}}^{(3)}$ interactions, respectively, as well as those of $U_{p_i p_{i+3}}^{el}$ and $U_{corr;\, p_i p_{i+3}}^{(3)}$, respectively, are handled by separate loops over the p_i, p_{i+2} and p_i, p_{i+3} pairs, respectively, to provide better load balance. The long-range electrostatic and correlation interactions are computed from the p_i, p_{i+4} pairs, onward. The loops to compute interactions are then distributed to the processors assigned to a CG task, subject to optimum load balancing. As mentioned earlier in this section, no cutoff is applied to the interactions listed in points 1c or 1d, and consequently, the lists of these interactions consigned to each fine-grained processor are fixed. For each energy term listed in points 1a and 1b, each fine-grained processor is assigned the range of the residues (in terms of the number of the residue to start and to end at) to compute this term. For the terms listed in points 1c and 1d, each processor is assigned a list of the start and end residue numbers to process for each first pair index (i). If the first element of entry i of the list of interactions that belongs to processor i is zero, this processor does not have any interaction such that residue i is the first in the pair to process. The number of components of each

of the energy terms that belongs to the 1a, 1b, or 1c category is optimally load-balanced for each processor and differs by 1 at most for different processors. For the interactions between the peptide groups, interaction distribution between processors (along with the assignment of the pertinent correlation interactions) is illustrated in Figure 4.

For the interactions listed in point 2, a 7 Å cutoff is applied to the distances between peptide groups ($r_{p_ip_j}$ and $r_{p_ip_{j\pm1}}$) involved in a correlation, with a contact function given by eq 42 of ref 29, which varies smoothly with the distance from 1 to 0 from $r = 6.9$ to $r = 7$ Å. Therefore, a list of interactions to be sent by a given processor to other processors and received by it from other processors is dynamic.

Before calculations are actually started, each processor (let its rank be R) creates the list of processors to which quantities pertaining to each pair of peptide groups might need to be sent and a list of processors from which it can receive terms pertaining to pairwise interactions. The algorithm to construct these lists is as follows:

Send List

(1) Loop through the p_ip_{i+2} pairs consigned to the processor. For each pair, loop through the other processors with ranks less than R, and add processor S to the send list of p_ip_{i+2} interaction at processor R if processor S has (i) the $p_{i-1}p_{i+1}$ interaction or (ii) the $p_{i-1}p_{i+3}$ interaction.

(2) Loop through the p_ip_{i+3} pairs consigned to the processor. For each pair, loop through the other processors with ranks less than R and add processor S to the send list of p_ip_{i+3} interaction at processor R if processor S has (i) the $p_{i-1}p_{i+2}$ interaction or (ii) the $p_{i-1}p_{i+4}$ interaction.

(3) Loop through all p_ip_j pairs consigned to the processor such that $j > i + 3$. For each pair, loop through the other processors with ranks less than R and add processor S to the send list of p_ip_j interaction at processor R if processor S has (i) the $p_{i-1}p_{j-1}$ interaction or (ii) the $p_{i-1}p_{j+1}$ interaction or (iii) the $p_{i+1}p_{j-1}$ interaction. Situation iii occurs only if $p_{i+1}p_{j-1}$ is a pair of 1,3- or 1,4-adjacent peptide groups (i.e., $j - i = 4$ or $j - i = 5$), because these interactions are distributed to processors independent of the longer-range interactions, and consequently, a processor with a rank lower than R might have a 1,3 or a 1,4 interaction with greater i than the long-range interactions consigned to processor R.

Receive List

The construction of the receive list of a receiving processor R is the reverse of the construction of the send list. Therefore, processors with ranks greater than R are searched in steps 1 and 3 by the receiving processor R, and if a processor is found to contain interactions that might be needed by R, its rank is added to the receive list of processor R. The ranks of the processors which R can potentially receive interactions from is the only information stored in the receive list, because the actual interaction list to be received from a given processor depends on conformation. The interaction list is, therefore, provided to R every time the correlation interactions consigned to this processor are calculated.

For processor R, each pair of peptide groups from the preset send lists is checked, and if the contact function value is greater than 0, all information pertaining to this peptide group pair, needed to compute correlation interactions and their gradients, is transferred to buffers intended for processors of the send list of this interaction. After the entire preset send list has been scanned, nonblocking send operations to each of the processors from the send list are initiated. If there are no interactions to be sent to processor S, the number 0 is sent to indicate this fact. After the send operations have been initiated, nonblocking receive operations are executed. A waitall operation completes the send−receive process to synchronize all fine-grained tasks. Nonblocking send and receive operations are used to exchange interaction information with a wait operation to synchronize the calculation of correlation energy. A scheme of information exchange is presented in Figure 4.

Each processor adds the received interactions to the list of its own computed interactions; however, it marks them as "received" to prevent double-counting of the respective correlation interactions. The latter could happen if, e.g., the processor has the interaction p_kp_l and needs to receive $p_{k+1}p_{l+1}$ and, at the same time, has $p_{k+1}p_{l-1}$ and needs to receive $p_{k+2}p_l$ to compute the respective correlation interactions. The received interactions $p_{k+1}p_{l+1}$ and $p_{k+2}p_l$ themselves form a correlation interaction (since a correlation interaction is formed both by p_ip_j and $p_{i+1}p_{j+1}$ and by p_ip_j and $p_{i+1}p_{j-1}$).[29] Consequently, if they were not marked "received", this additional interaction would be computed by the processor that received both of them, and since this interaction is computed by the processor which owns $p_{k+1}p_{l+1}$, the interaction would be doubly counted. It should be noted that marking the received interactions does not imply any additional communication. A correlation term pertaining to p_kp_l and $p_kp_{l\pm1}$ is not computed if both pairs of interactions are marked "received". This is because a given processor always has one interaction of its own of a pair of interactions that constitute a correlation term (see Figure 4).

As described above, the distribution of correlation interactions between processors provides a nearly optimal load balance, given the fact that the interactions between peptide groups are optimally distributed between the fine-grained processors. If the 7 Å cutoff, which is imposed on each of the pairs of interacting residues to compute the respective correlation interaction,[29] is not in effect (for small proteins), the number of correlation terms is twice the number of interactions consigned to the processor except when it has the first or the last residue. For medium-sized and large proteins, the cutoff which decreases the number of correlations is in effect; however, even in this case, the number of correlation interactions is nearly the same for each fine-grained processor, because the number of peptide groups within the cutoff distance from a given peptide group remains proportional to the total number of peptide groups with which it interacts.

2.5.4. Fine-Grained Parallelization of Other Operations. Computing accelerations from forces in UNRES involves multiplication of the vector of forces expressed in virtual-bond vectors by the inverse of the matrix G [eq 12]. Since the part of the forces is expressed in internal coordinates or in Cartesian coordinates of the UNRES interaction sites, these forces first need to be transformed to the space of

virtual-bond vectors. U_b, U_{tor}, and U_{tord}, as well as the correlation terms except $U_{corr}^{(4)}$ are expressed in backbone virtual-bond (θ) and virtual-bond-dihedral (γ) angles. The transformation takes the following form:

$$\nabla'_{\mathbf{dC}_i} U_X = \frac{\partial U_X}{\partial \theta_{i-1}} \nabla_{\mathbf{dC}_i} \theta_{i-1} + \frac{\partial U_X}{\partial \theta_i} \nabla_{\mathbf{dC}_i} \theta_i + \frac{\partial U_X}{\partial \gamma_{i-2}} \nabla_{\mathbf{dC}_i} \gamma_{i-2} + \frac{\partial U_X}{\partial \gamma_{i-1}} \nabla_{\mathbf{dC}_i} \gamma_{i-1} + \frac{\partial U_X}{\partial \gamma_i} \nabla_{\mathbf{dC}_i} \gamma_i \quad (14)$$

where U_X is U_b, U_{tor} or U_{tord}, $U_{corr}^{(3)}$, $U_{corr}^{(5)}$, $U_{corr}^{(6)}$, U_{turn}^{3}, $U_{turn}^{(4)}$, and $U_{turn}^{(6)}$ and $\nabla'_{\mathbf{dC}}$ indicates that we consider only the part of the derivatives corresponding to dependence on the θ and γ angles (it should be noted that only U_b depends on θ angles). Therefore, the transformation of this part of the gradient involves an $\mathcal{O}(n)$ effort, where n is the number of residues and can be predicted not to benefit from parallelization. Nevertheless, we included an option to fine-grain the operations given by eq 14, but communication overhead was greater than the profit from fine-graining, except for a small number of processors. The rotamer potentials (U_{rot}) and the virtual-bond-deformation potentials (U_{bond}) are expressed directly as functions of virtual-bond vectors,[38] and consequently, no force transformation is needed for these terms.

The derivatives of both the pairwise and correlation terms in $\mathbf{q}$ are expressed by eq 15.

$$\nabla_{\mathbf{q}} U_X = \mathbf{A}^{\mathbf{T}} \nabla_{\mathbf{x}} U_X \quad (15)$$

where $\nabla_{\mathbf{y}} = (\partial/\partial y_1, \partial/\partial y_2, ..., \partial/\partial y_n)$ is the gradient operator. Recalling eq 6, it can be realized that the transformation of the energy gradient in site positions ($\mathbf{x}$) to that in the virtual-bond vectors forming the vector of generalized coordinates $\mathbf{q}$ [eq 3] involves only summations [the effort being roughly $\mathcal{O}(n^2)$] and divisions by 2 [the effort being $\mathcal{O}(n)$]. This observation was implemented in writing the code. Nevertheless, the $\mathcal{O}(n^2)$ effort requires parallelizing this part of the computations.

The matrix $\mathbf{A}$ and the vector $\nabla_{\mathbf{x}} U_X$, as well as the matrix $\mathbf{G}$ and vector $\mathbf{F}$, are divided between processors, as shown in Figure 5. Each processor executes its part of the computations, and reduction to the master fine-grain task with the sum operator (in the MPI terms) is executed at the end.

3. Results and Discussion

3.1. Fine-Graining a Single MD Trajectory. To assess the performance of the fine-grained part of the code, we carried out short canonical MD simulations with the Berendsen thermostat[57] implemented in UNRES[42] for the following nine proteins with size from 37 to 767 residues: 1E0L (37 residues), 1BDD (46 residues), 1KOY (62 residues), 2K4N (111 residues), 2K5I (154 residues), 1P1D (196 residues), 3G5A (304 residues), 2KHO (600 residues), and 1TF5 (767 residues). We used the recent force field calibrated by extensive random exploration of energy-parameter space.[56] The tests were run on the machines listed in section 2.4. From 100 to 10 000 UNRES MD steps, depending on protein size and processor speed, with

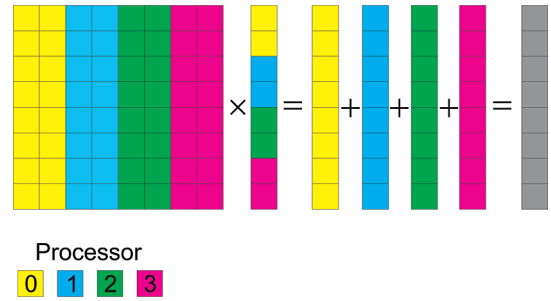


Figure 5. Partitioning of the matrix **A** and vector $\nabla_{\mathbf{x}} U$(a) and of the matrix **G** and the vector $\nabla_{\mathbf{q}}$ between the fine-grain tasks for distributed computing of $\nabla_{\mathbf{q}}$ and accelerations, respectively. Different colors mark different processors (rank shown in the small lower panel). The matrix−vector multiplication is distributed between processors (each handling only its part of the matrix and of the vector) to give the vectors containing incomplete sums in each component. The vectors are subsequently summed up to give the resulting vector.

1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024 processors, respectively, were taken; the largest number of processors was tested only on the machines with the fastest communication (jugene.fz-juelich.de and intrepid.alcf.anl.gov). The number of MD steps was the same for a given protein and machine (the strong-scaling runs).

The nonsetup time (defined as the wall-clock time used to perform MD steps) was taken to assess parallel performance. The setup time, not taken into consideration, includes data reading, computing and inverting the inertia tensor, and generating initial velocities. For production runs, the setup time is negligible. The nonsetup times per MD step for the above-mentioned proteins, each run with a single processor of the machines listed above, are plotted in Figure 6a; additionally, the per-day simulation lengths (in nanoseconds/day), together with the per-day lengths corresponding to the all-atom GROMACS 4.0.5 program run on jugene.fz-juelich.de for 1BDD, 1P1D, and 1TF5, are plotted in Figure 6b. As can be seen from this figure, in all cases the single-processor nonsetup time grows with the square of the number of residues (a slope of about 2 in all four plots of Figure 6), as expected because most of the computational effort is $\mathcal{O}(n^2)$, where n is the number of residues in a polypeptide chain. For a given protein, the non-setup times increase in the order galera.task.gda.pl < bigben.psc.edu < jugene.fz-juelich.de < intrepid.alcf.anl.gov, which conforms to the processor speed of these machines. With cutoff introduction on nonbonded interactions, the GROMACS computation time scales almost linearly with protein size.

In Figure 7a−d and Figure 8a−d the speedups and efficiencies, respectively, for the nine proteins obtained with galera.task.gda.pl, bigben.psc.edu, intrepid.alcf.anl.gov, and jugene.fz-juelich.de, respectively, are plotted as functions of the numbers of processors. The speedup (s) and efficiency (η) are calculated from eqs 16 and 17, respectively.

$$s(P) = \frac{t(1)}{t(P)} \quad (16)$$

$$\eta(P) = \frac{t(1)}{Pt(P)} \tag{17}$$

where P is the number of processors, $t(1)$ is the time of a single-processor run, and $t(P)$ is the time of a run with P processors.

It can be seen that the best scalability is achieved with both IBM BlueGene/P supercomputers (a speedup over 4 is achieved even for the smallest protein 1E0L), while galera.task.gda.pl represents the poorest parallel performance except for runs for the smallest proteins with up to 4 processors (in which case the quad-core architecture of this computer is an advantage) for which it has better performance than bigben.psc.edu. These observations conform to the difference in the computation and communication speed of the four supercomputers mentioned. As an example, the communication times pertaining to collective operations (MPI_REDUCE, MPI_GATHER, and MPI_SCATTER) expressed relative to single-processor time and relative to the wall-clock time of a given multiprocessor run are plotted in Figure 9a and b, respectively, for the largest protein (1TF5; 767 residues) as function of the number of processors. As can be seen from Figure 9, the ratio of the communication time is the smallest for both IBM BlueGene/P machines; moreover, it varies little between 8 and 128 processors. With machines with not so fast communication, the communication time constitutes even 70% of wall-clock time. Because of a nearly constant communication-time overhead, the speedup and efficiency obtained from the runs on BlueGene/P computers for larger proteins quite strictly obey the classical Amdahl law [eq 18].[66]

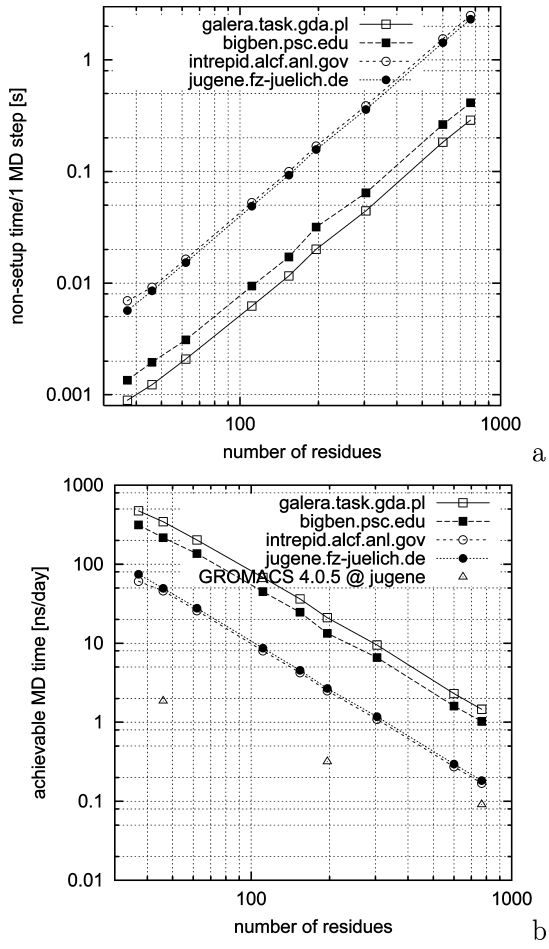


Figure 6. Plots of (a) nonsetup times per MD step and (b) the achievable simulation length (ns/day) as functions of number of residues with a single processor for the four supercomputers used in this study. A logarithmic scale is used on both axes; it can be seen that the slope is approximately 2, indicating a quadratic dependence on the number of residues. The per-day simulation lengths achievable with GROMACS 4.0.5 at jugene.fz-juelich.de are also included for comparison. The nonsetup time per MD step (in seconds) can be expressed approximately as $t = 4.95 \times 10^{-7}n^2$ for galera.task.gda.pl, $t = 7.0 \times 10^{-7}n^2$ for bigben.psc.edu, $t = 4.26 \times 10^{-6}n^2$ for intrepid.alcf.anl.gov, and $t = 3.94 \times 10^{-6}n^2$ for jugene.fz-juelich.de, respectively, where n is the number of residues.

$$s(P) = \left[f + \frac{1-f}{P}\right]^{-1} \tag{18}$$

where s is the speedup, P is the number of the processors, and f is the fraction of time spent when running the nonparallelized part of the code with one processor. The dependence of the communication time on the number of processors exhibits the largest slope for galera.task.gda.pl.

A more detailed analysis of the timing is provided in Table 1 with the example of 1TF5 (the largest protein considered in our study) run on bigben.psc.edu. It can be seen that, although the nonparallelized operations (such as calculations of virtual-bond angles and virtual-bond dihedral angles and calculation of auxiliary quantities to compute $U_{corr}^{(3)}$ and $U_{turn}^{(3)}$ and $U_{turn}^{(4)}$) constitute a negligible fraction of the CPU time with a single processor, they take relatively more and more time with an increasing number of processors and, consequently, reduce the available speedup according to Amdahl's law and gradually become the bottleneck of the computations. As pointed out in section 2.5.4, we attempted to parallelize these operations, but the communication overhead turned out to be greater than the gain from eliminating this nonparallelized component of computations. However, it is possible that the scalability of UNRES calculations can benefit from parallelizing the above-mentioned operations on mixed shared-and-distributed memory architectures. Additionally, the calculation of $U_{corr}^{(4)}$ is not fully load-balanced at present (see section 2.5.3) and, consequently, does not scale very well (see Table 1); however, the relative contribution of the computation of $U_{corr}^{(4)}$ to the total computation time is small. Nevertheless, a better parallelization of this energy component is necessary for further extension of the time scale of UNRES calculations. The load imbalance can be addressed in the force-decomposition scheme by randomizing distribution of peptide groups on processors in a single preprocessing step.[63,64]

A practical issue in molecular simulations is the wall-clock time in which simulations can be accomplished. In Figure 10a, the minimum nonsetup times per MD (irrespective of efficiency of a parallel run) and, in Figure 10b, the nonsetup times at 50% efficiency are plotted vs the number of residues in a chain. Additionally, the data corresponding to 50% efficiency are presented in Figure 10c as the achievable length of simulation expressed in nanoseconds of MD

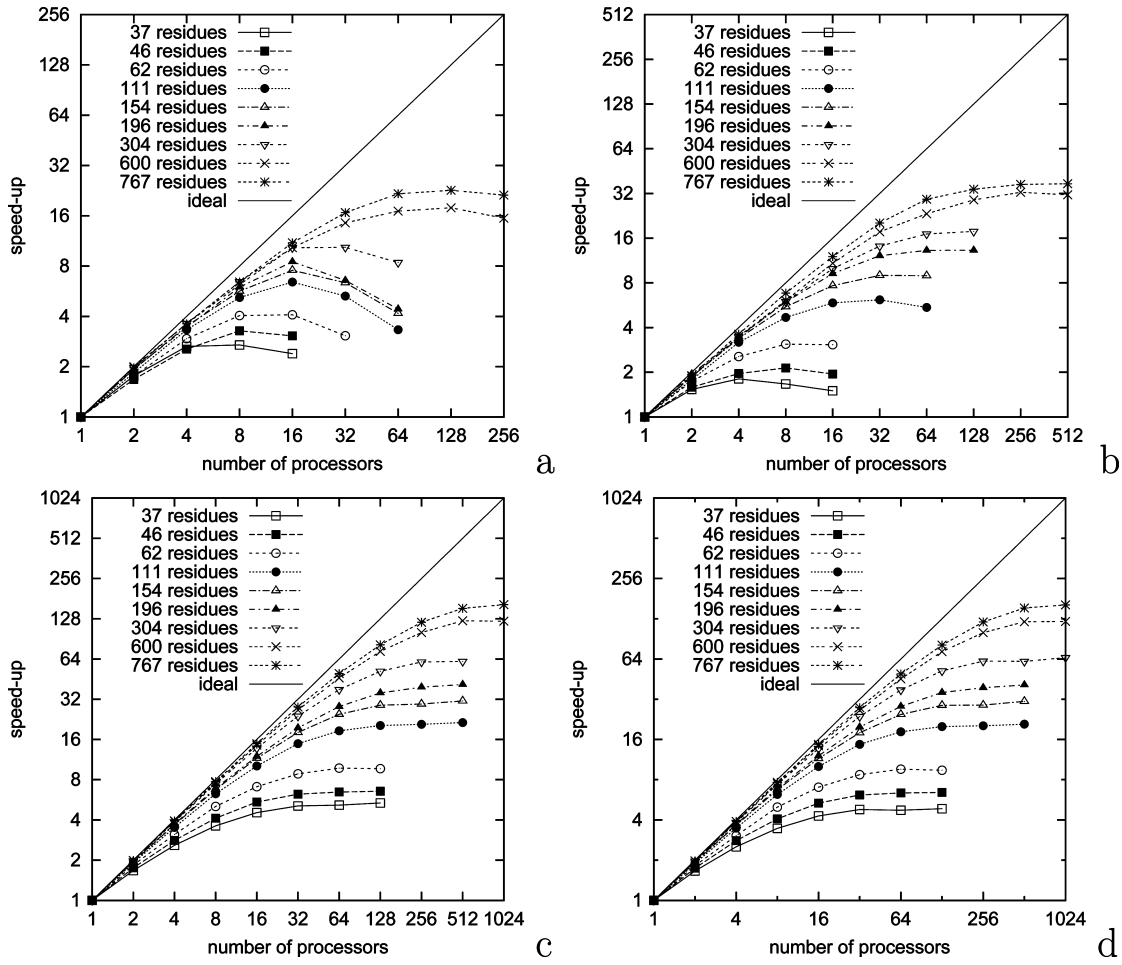


Figure 7. Plots of the speedups for canonical single-trajectory UNRES/MD runs obtained for proteins with various numbers of residues vs the number of processors on (a) galera.task.gda.pl, (b) bigben.psc.edu, (c) intrepid.alcf.anl.gov, and (d) jugene.fz-juelich.de. The logarithmic scale with base 2 is used on both axes.

simulation per day, which is a commonly used metric.[15] It can be seen that the curves presented in Figure 10a and b are slightly convex initially, then approximately linear, and for the largest proteins, the slope starts to decrease. The convex shape of the initial portion of the time vs number of residues plots is the most pronounced for galera.task.gda.pl, while for IBM BlueGene/P supercomputers, the dependence of the minimum nonsetup time and the nonsetup time at 50% efficiency on the number of residues is nearly linear. The approximately linear dependence of the wall-clock time per MD step on protein size is a big advantage over single-processor runs in which this dependence is quadratic (Figure 6). On the other hand, it can be seen from Figure 10 that fast communication does not overcome greater processor speed. The lines corresponding to both IBM BlueGene/P supercomputers are always above those of the other two machines with faster processors. For the 1TF5 protein (the largest one), the minimum and the 50% efficiency times are longer by about 12% and 30% on IBM BlueGene/P than on galera.task.gda.pl and bigben.psc.edu, respectively. On the other hand, this is still better than the ratio of the single-processor time of the faster IBM BlueGene/P (jugene.fz-juelich.de) to that of galera.task.gda.pl and bigben.psc.edu, which are 6.36 and 4.21, respectively (however, 4 times more processors have to be used on IBM BlueGene/P).

For reference, the single-processor nonsetup times of the small 1KOY protein (62 residues) are also indicated in Figure 10. For this small protein, 20 000 000 MD steps (100 ns with the 5 fs time step used in UNRES simulations), which is required for a converged canonical or replica-exchange UNRES/MD run, can be accomplished in about 12 h of wall-clock time on galera.task.gda.pl and in about 18 h on bigben.psc.edu, respectively, without fine-graining the code. It can be seen from Figure 10 that fine-graining enables us to accomplish simulations for about 200-residue proteins on galera.task.gda.pl and bigben.psc.edu in the same wall-clock time as for 1KOY. For the two IBM BlueGene/P computers, simulations of the largest (1TF5) protein can be accomplished in a shorter wall-clock time than that required for 1KOY with these machines, if efficiency is neglected (Figure 10a), or in a slightly greater wall-clock time with 50% efficiency (Figure 10b). With the maximum speedup achievable, 20 000 000 MD-step simulations for 1TF5 can be accomplished in about 80 wall-clock hours (3.3 wall-clock days). A millisecond of simulations for this protein (effectively a second, given the ~1000-times lengthening of the UNRES time scale with respect to the experimental time scale)[42,43] would require only slightly longer than a month of computations on IBM BlueGene/P. This achieve-

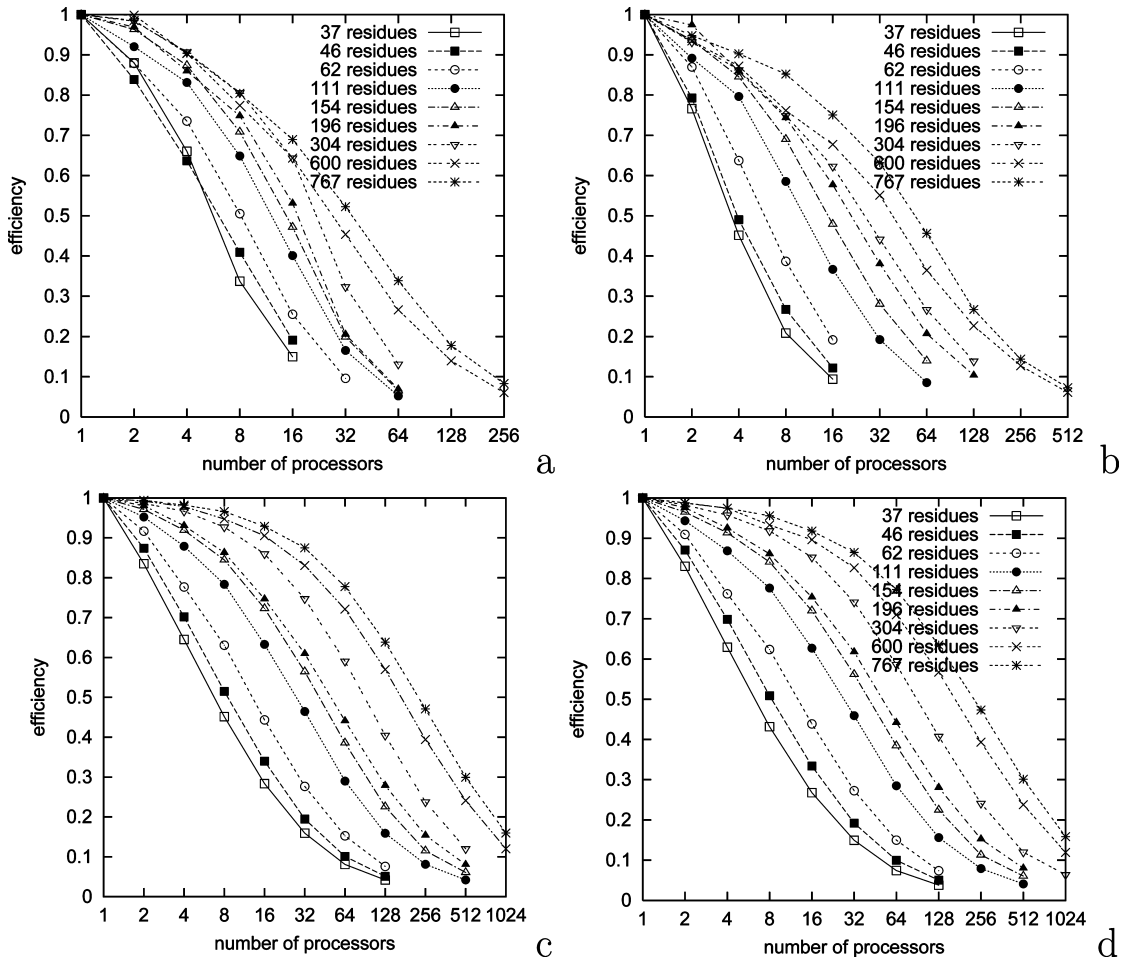


Figure 8. Plots of the efficiencies for canonical single-trajectory UNRES/MD runs obtained for proteins with various numbers of residues vs the number of processors on (a) galera.task.gda.pl, (b) bigben.psc.edu, (c) intrepid.alcf.anl.gov, and (d) jugene.fz-juelich.de. The logarithmic scale with base 2 is used on the abscissae.

ment makes the biological time scales achievable with the UNRES model.

In Figure 10c, the data corresponding to all-atom calculations with GROMACS[15] at 50% efficiency with jugene.fz-juelich.de are also presented. GROMACS is currently one of the most efficient and best load-balanced MD software programs. An 11 Å cutoff was applied on all nonbonded interactions. As shown, UNRES provides 5−7 times longer per-day simulation length compared to GROMACS. Because the UNRES event-based time scale is at least 4 times longer than that of the all-atom approach,[42,43] the effective speedup with respect to GROMACS is at least 20 times. When the single-processor times are compared, the UNRES per-day simulation length is 27 times greater for 1BDD, 8 times greater for 1P1D, and 2 times greater for 1TF5 (see also Figure 6b). The decrease of the UNRES/GROMACS per-day simulation length with a single processor is not surprising, because UNRES does not implement a cutoff on nonbonded interactions, as opposed to GROMACS and, consequently, the complexity of UNRES is $\mathcal{O}(n^2)$, while that of GROMACS (which implements the cutoff) is $\mathcal{O}(n)$. Therefore, introducing a cutoff on nonbonded interactions in UNRES and, thereby, reducing the complexity to $\mathcal{O}(n)$, could be advantageous for large proteins. Preliminary results with an 11 Å cutoff on nonbonded interactions in UNRES have shown that the scaling becomes almost linear and the ratio of single-processor per-day simulation time with UNRES to that with GROMACS becomes 10 for 1TF5.

3.2. Two-Grain Multiplexed Replica-Exchange Simulations. The scalability data reported in section 3.1 pertain to single-trajectory runs or canonical runs with multiple independent trajectories. In the replica-exchange (REMD) and multiplexed replica exchange molecular dynamics (MREMD) simulations, exchange of information occurs. Fine graining might influence the scalability of the coarse-grained tasks because even residual load imbalance can cause remarkable differences in the wall-clock time required to compute the number of MD steps between exchanges. In this section, we, therefore, report the results of the scalability of short MREMD runs. We chose only three proteins: 1KOY (small size), 1P1D (medium size), and 2KHO (large size). The calculations were run on galera.task.gda.pl, bigben.psc.edu, and jugene.fz-juelich.de; we omitted intrepid.alcf.anl.gov because it gives the same scalability profiles as those of jugene.fz-juelich.de. We ran from 1 to 256 MREMD trajectories (the numbers of trajectories were the consecutive powers of 2 and simulations for 1 trajectory were taken as reference because they were canonical MD simulations). The number of fine-grained processors was generally either 1 (reference for an MREMD simulation with a given number

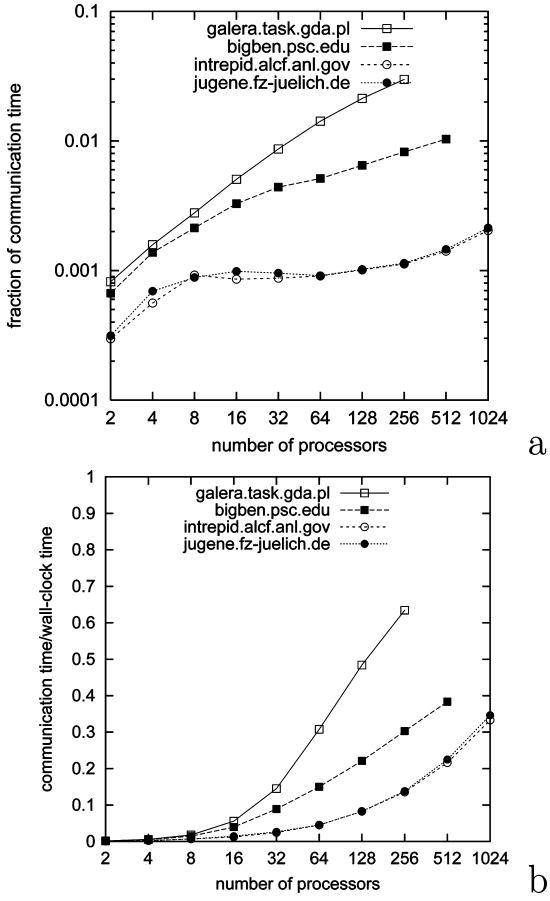


Figure 9. Plots of (a) the fraction of the collective-communication time to single-processor nonsetup time vs the number of processors in a single-trajectory UNRES/MD run for 1TF5 and (b) fraction of the collective-communication time to the wall-clock time with a given number of processors. Logarithmic scales are used on both axes.

of trajectories) or corresponding to about 50% efficiency for a given protein and a given machine (section 3.1). However, in runs on the IBM BlueGene/P machine, we also used the number of fine-grained processors corresponding to nearly maximum scalability. The number of MD steps per trajectory and the frequency of replica exchange were the same for a given protein (weak-scaling tests); 40 000 steps with exchange each 10 000 steps for 1KOY, 20 000 steps with exchange each 5000 steps for 1P1D, and 10 000 steps with exchange each 5000 steps for 2KHO, respectively. The most massively parallel calculation was that run for 2KHO on jugene.fz-juelich.de; it comprised 256 trajectories (CG tasks), each run with 256 processors (FG tasks), i.e., 65 536 processors total.

Plots of the nonsetup time and efficiency vs the number of trajectories (nonfine grained tasks) for different numbers of fine-grain tasks are shown in Figures 11, 12, and 13 for 1KOY, 1P1D, and 2KHO, respectively. It can be seen that the efficiency of MREMD calculations decreases slightly with the number of trajectories, although the decrease is not remarkable. The most remarkable decrease is observed for the runs on the IBM BlueGene/P supercomputer; there is a clear efficiency drop (about 5%), after which the efficiency does not decrease remarkably with the number of trajectories. It can be noted that, except for the smallest 1KOY protein,

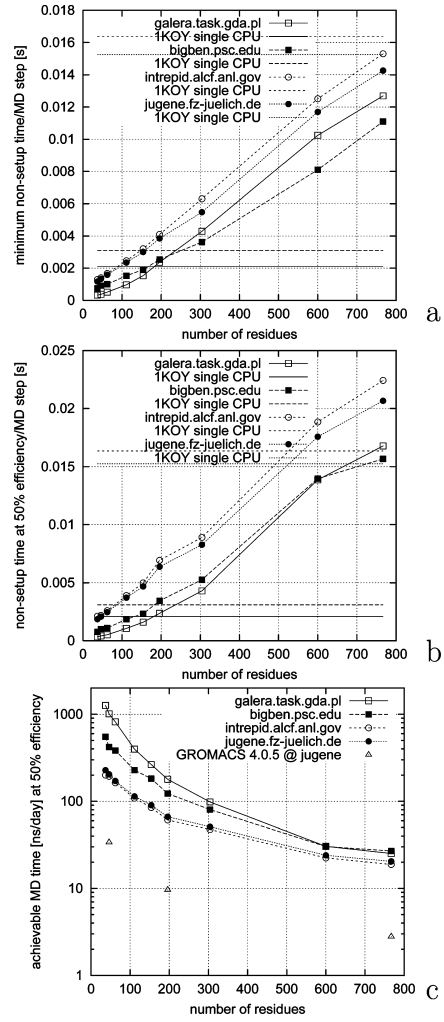


Figure 10. Plots of the minimum nonsetup time per MD step (a), the nonsetup time per MD step with 50% parallel efficiency (b), and the achievable formal simulation length (in nanoseconds/day; taking the 4.89 fs MD time step) with 50% efficiency (c) vs the number of residues obtained in fine-grained single-trajectory UNRES/MD runs on galera.task.gda.pl, bigben.psc.edu, intrepid.alcf.anl.gov, and jugene.fz-juelich.de. For references, horizontal lines (solid for galera.task.gda.pl, long-dashed for bigben.psc.edu, short-dashed for intrepid.alcf.anl.gov, and dotted for jugene.fz-juelich.de) are drawn which correspond to single-processor nonsetup times of the 62-residue 1KOY protein. The time at 50% efficiency was obtained by linear interpolation from the times at the efficiencies bracketing 50%. The achievable simulation lengths at 50% efficiency obtained for 1BDD, 1P1D, and 1TF5 proteins with GROMACS 4.0.5[15] (taking the 2 fs time step) run at jugene.fz-juelich.de are also included in part c for comparison. For 1E0L, 1BDD, 1KOY, 2K4N, 2K5I, 1P1D, 3G5A, 2KHO, and 1TF5, respectively, the numbers of processors corresponding to part a are 8, 8, 16, 16, 16, 16, 32, 128, and 128 with galera.task.gda.pl; 4, 8, 8, 32, 32, 128, 128, 256, and 512 with bigben.psc.edu; and 128, 128, 64, 128, 128, 512, 1024, 1024, and 1024 with intrepid.alcf.anl.gov and jugene.fz-juelich.de. The numbers of processors corresponding to UNRES runs reported in parts b and c are 8, 8, 16, 16, 16, 16, 32, 128, and 128 with galera.task.gda.pl; 4, 8, 8, 32, 32, 128, 128, 256, and 512 with bigben.psc.edu; and 8, 8, 16, 32, 32, 64, 64, 128, and 256 with intrepid.alcf.anl.gov and jugene.fz-juelich.de. The numbers of processors used in the GROMACS runs reported in part c were 32 for 1BDD and 64 for 1P1D and 1TF5.

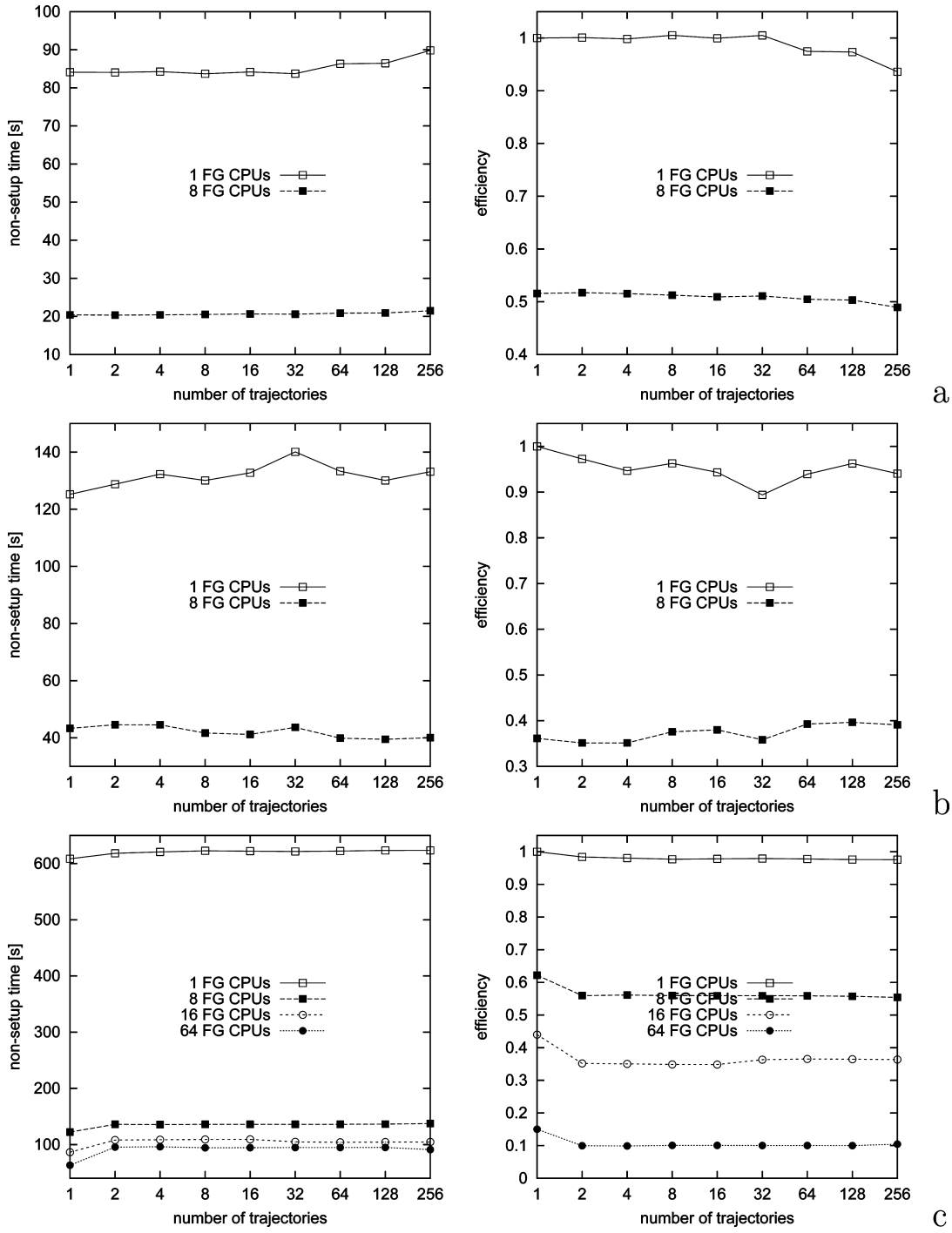


Figure 11. Plots of nonsetup times and efficiencies in 40 000-step MREMD simulations with replica exchange each 10 000 steps, numbers of trajectories from 1 (reference) to 256, and various numbers of fine-grain tasks per trajectory obtained for the 1KOY (62-residue) protein with (a) galera.task.gda.pl, (b) bigben.psc.edu, and (c) jugene.fz-juelich.de.

the drop of efficiency or the increase of the nonsetup time for IBM BlueGene/P is nearly constant. By detailed timing of the runs, we found that the decrease of efficiency is caused by increasing waiting times by the master processors which distributes coarse-grained tasks (Figure 2). Most probably, this waiting time increases because the calculations for each trajectory take slightly different times, which results in a load imbalance of the coarse-grained tasks.

4. Conclusions

In this work, we have extended the parallelization of UNRES MD from the already-existing one processor per trajectory mode,[51] by fine-graining single-trajectory calculations. Even for the smallest protein (1E0L, 37 residues), a speedup from about 2 to about 5 can be achieved depending on the machine (the largest on IBM BlueGene/P which has the fastest communication); for the largest protein studied (1TF5, 767

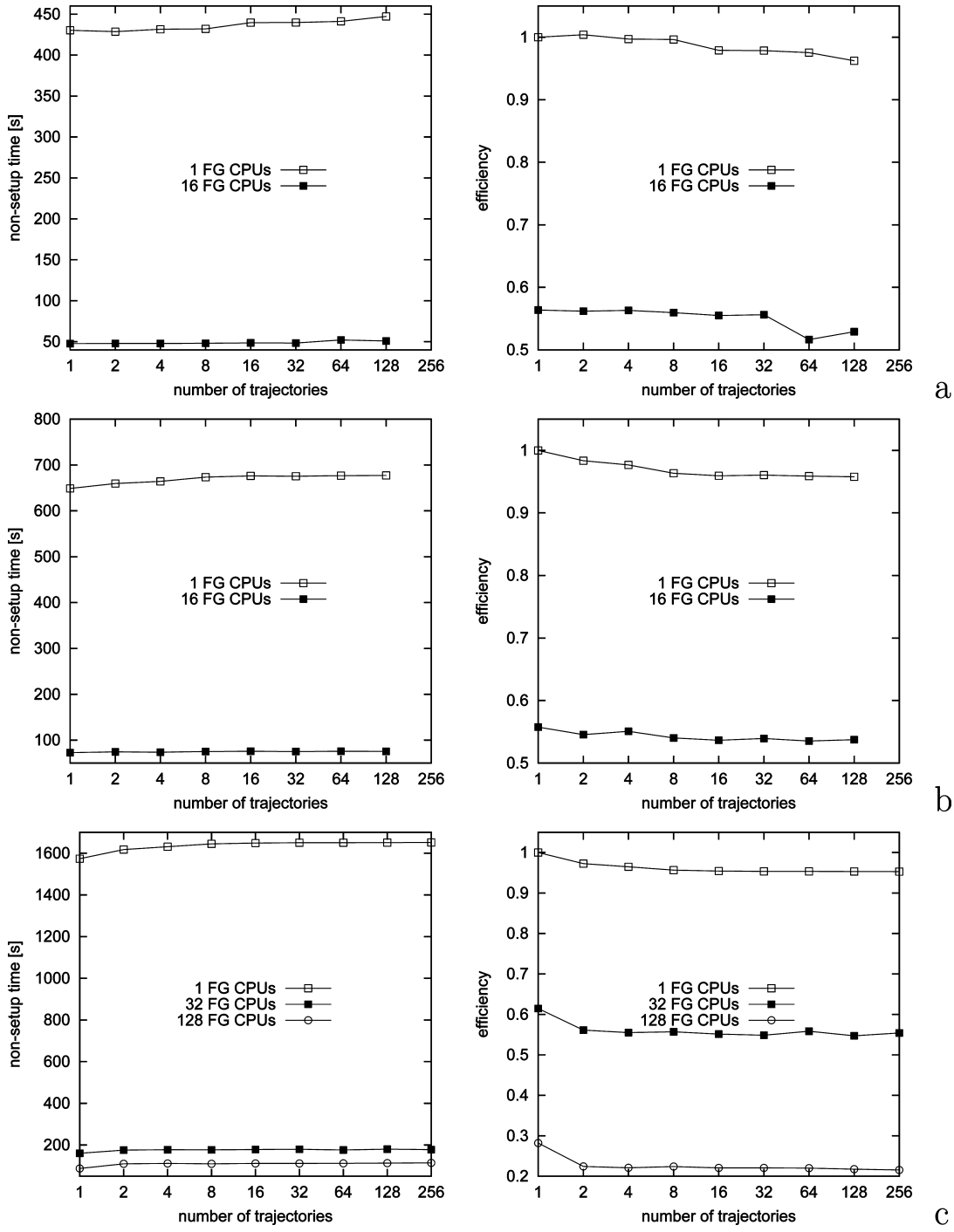


Figure 12. Plots of nonsetup times and efficiencies in 20 000-step MREMD simulations with replica exchange each 5000 steps, numbers of trajectories from 1 (reference) to 256, and various numbers of fine-grain tasks per trajectory obtained for the 1P1D (196-residue) protein with (a) galera.task.gda.pl, (b) bigben.psc.edu, and (c) jugene.fz-juelich.de.

residues), a 160-fold maximum speedup has been reached with 1024 processors/trajectory, and 120-fold speedup at 50% parallel efficiency with 256 processors/trajectory was reached with IBM BlueGene/P. Even for machines with slower communications (but 4- or 6-times greater processor speed than that of IBM BlueGene/P), a reasonable speedup of 32 or 16 can be achieved at 50% efficiency with the Cray XT3 and Xeon cluster, respectively. The efficiency is slightly diminished for MREMD runs which involve communication between coarse-grained tasks; however, this does not seem to be a significant issue. Given the fine-grained parallelism, the wall-clock time necessary to compute a single trajectory increases linearly with the number of residues, as opposed to an increase with the square of the number of residues when using a single processor per trajectory. For the 1TF5 protein, a single MD step can be accomplished in 0.015 wall-clock s, which means that a 20 000 000 step run (0.1 μs with 5 fs MD time step) can be accomplished in about 3.5 days of simulations. Given the fact that the UNRES (and, generally, a coarse-grained approach) time scale is about 1000 times wider than that of the experimental time scale (i.e., protein folding with UNRES takes only nanoseconds,[42,43] while the

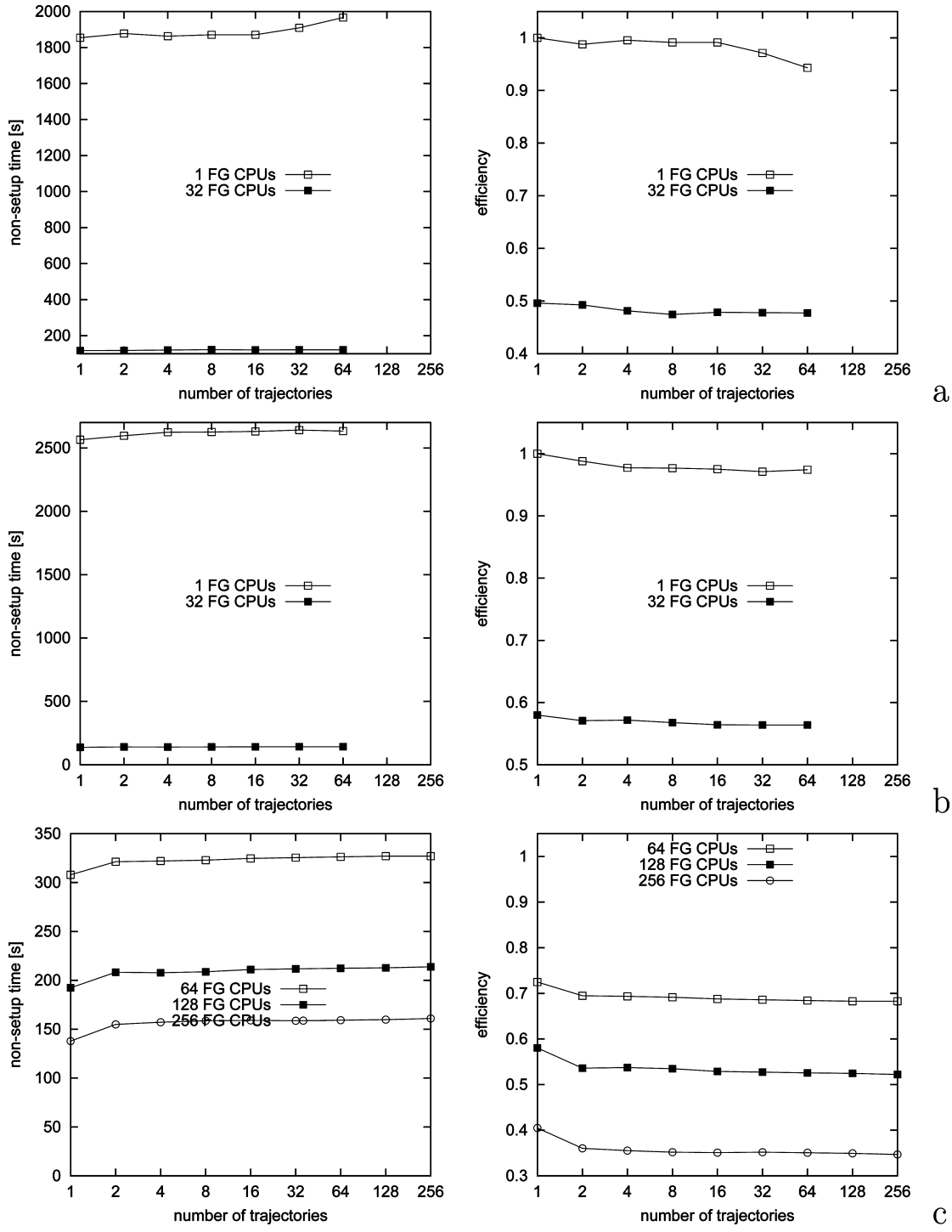


Figure 13. Plots of nonsetup times and efficiencies in 10 000-step MREMD simulations with replica exchange each 5000 steps, numbers of trajectories from 1 (reference) to 256, and various numbers of fine-grain tasks per trajectory obtained for the 2KHO (600-residue) protein with (a) galera.task.gda.pl, (b) bigben.psc.edu, and (c) jugene.fz-juelich.de. The reference time (1 trajectory with 1 fine-grained processor) corresponding to the runs of part c was calculated from the 2KHO entry in the data plotted in Figure 6.

fastest-folding proteins fold in microseconds),[44] this achievement enables us to carry out millisecond-scale simulations of large-size proteins in real time. Introducing a cutoff on nonbonded interactions, as pointed out in section 3.1, might push the achievable time scale even farther or, at least, reduce the cost of calculation, and we are currently working on this modification. As mentioned in section 3.1, the load balance can be achieved with a cutoff in the force-decomposition scheme;[63,64] however, we will also explore the domain-decomposition scheme.

At present, UNRES requires about 1 GB/processor memory to run calculations for ~900-residue proteins. A large part of this memory is occupied by matrix **G**, its inverse, and auxiliary matrices. This poses some problems, although a great part of this could be eliminated by switching to single precision (which, additionally, should reduce the CPU time)

and by putting the arrays in shared memory. Because the array **G** for multichain proteins consists of independent blocks, each pertaining to a different chain, and the maximum number of residues per chain usually does not exceed 2000, this modification will solve a large part of the memory problem. The second substantial part of memory is required to store the information to compute the $U_{corr}^{(4)}$ terms; however this part can be distributed between fine-grained processors. This work is presently being carried out in our laboratory.

The successful fine-grain parallel implementation of UNRES reported in this work suggests that the UNRES code can be ported efficiently to the GPUs, reaching an even greater speedup of the fine-grained part of the code than on IBM BlueGene/P; for all-atom MD, a 700-fold speedup was recently reported.[18] In particular, large constant arrays belonging to given conformations could be shared by all processors of a GPU unit, which would eliminate the present problem occurring with distributed-memory fine-grained UNRES. However, even though implementation of UNRES on the GPUs seems to be very attractive, it is not clear if a similar speedup can be achieved with UNRES as that with all-atom-code implementation on GPUs. We are currently in the process of porting UNRES code to GPUs.

Acknowledgment. This work was supported by grants from the Polish Ministry of Science and Higher Education (N N204 049035), the National Institutes of Health (GM-14312), and the National Science Foundation (MCB05-41633). This research was supported in part by the National Science Foundation through TeraGrid resources provided by Pittsburgh Supercomputing Center. Computational resources were also provided by (a) Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357, (b) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Jülich, Germany, and (c) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk. The following local resources—(a) the Beowulf cluster at the Department of Computer Science, Cornell University, (b) our 800-processor Beowulf cluster at Baker Laboratory of Chemistry, Cornell University, and (c) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk—were used to develop and test the code.

References

(1) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.

(2) Tirado-Rives, J.; Jorgensen, W. L. Molecular dynamics simulations of the unfolding of an α-helical analogue of ribonuclease A S-peptide in water. *Biochemistry* **1991**, *30*, 3864–3871.

(3) Daggett, V.; Levitt, M. Molecular dynamics simulations of helix denaturation. *J. Mol. Biol.* **1992**, *223*, 1121–1138.

(4) Brooks III, C. L. Characterization of native apomyoglobin by molecular dynamics simulations. *J. Mol. Biol.* **1992**, *227*, 375–380.

(5) Mark, A. E.; van Gunsteren, W. F. Simulation of the thermal-denaturation of hen egg-white lysozyme: trapping the molten globule state. *Biochemistry* **1992**, *31*, 7745–7748.

(6) Tirado-Rives, J.; Jorgensen, W. L. Molecular dynamics simulations of the unfolding of apomyoglobin in water. *Biochemistry* **1993**, *32*, 4175–4184.

(7) Daggett, V.; Levitt, M. Protein unfolding pathways explored through molecular-dynamics simulations. *J. Mol. Biol.* **1993**, *232*, 600–619.

(8) Day, R.; Daggett, V. All-atom simulations of protein folding and unfolding. *Adv. Protein Chem.* **2003**, *66*, 373–403.

(9) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. How well can simulation predict protein folding kinetics and thermodynamics. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43–69.

(10) Adcock, S. A.; McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(11) Daggett, V. Protein folding simulations. *Chem. Rev.* **2006**, *106*, 1898–1916.

(12) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.

(13) Kmiecik, S.; Kolinski, A. Folding pathway of the B1 domain of protein G explored by multiscale modeling, *Biophys. J.* **2008**, *94*, 726–736.

(14) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S.; Kaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. Atomistic protein folding simulations on the submillisecond timescale using worldwide distributed computing. *Biopolymers* **2003**, *68*, 91–109.

(15) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(16) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(17) 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. *ACM/IEEE SC 2006 Conference (SC.06)* **2006**, 43.

(18) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* **2009**, *30*, 864–872.

(19) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91–97.

(20) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(21) Kolinski, A.; Skolnick, J. Reduced models of proteins and their applications. *Polymer* **2004**, *45*, 511–524.

(22) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(23) Colombo, G.; Micheletti, C. Protein folding simulations: combining coarse-grained models and all-atom molecular dynamics. *Theor. Chem. Acc.* **2006**, *116*, 75–86.

(24) Ayton, G. S.; Noid, W. G.; Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(25) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

(26) Voth, G. Introduction. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1st ed.; Voth, G., Ed; CRC Press, Taylor & Francis: Boca Raton, FL, 2008; pp 1−4.

(27) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997**, *18*, 849–873.

(28) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J. Comput. Chem.* **1997**, *18*, 874–887.

(29) Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Ołdziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. United-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J. Comput. Chem.* **1998**, *19*, 259–276.

(30) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* **2001**, *115*, 2323–2347.

(31) Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H. A. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1937–1942.

(32) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. Determination of the potentials of mean force for rotation about $\mathbf{C}^{\alpha}\cdots\mathbf{C}^{\alpha}$ virtual bonds in polypeptides from the *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline. *J. Phys. Chem. A* **2003**, *107*, 8035–8046.

(33) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. Parameterization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from *ab initio* energy surfaces of model systems. *J. Phys. Chem. B* **2004**, *108*, 9421–9438.

(34) Ołdziej, S.; Łagiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization. *J. Phys. Chem. B* **2004**, *108*, 16950–16959.

(35) Kozłowska, U.; Liwo, A.; Scheraga, H. A. Determination of virtual-bond-angle potentials of mean force for coarse-grained simulations of protein structure and folding from ab initio energy surfaces of terminally-blocked glycine, alanine, and proline. *J. Phys.: Condens. Matter* **2007**, *19*, 285203.

(36) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **2007**, *111*, 260–285.

(37) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Rojas, A. V.; Kaźmierkiewicz, R.; Makowski, M.; Murarka, R. K.; Scheraga, H. A. Simulation of protein structure and dynamics with the coarse-grained UNRES force field. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1st ed.; Voth, G., Ed; CRC Press, Taylor & Francis: Boca Raton, FL, 2008; pp 1391−1411.

(38) Kozłowska, U.; Liwo, A.; Scheraga, H. A. Determination of side-chain-rotamer and virtual-bond-stretching potentials of mean force for coarse-grained simulations of protein structure and folding from AM1 energy surfaces of terminally-blocked amino-acid residues. 1. The Method. *J. Comput. Chem.* **2010**, in press; DOI: 10.1002/jcc.21399.

(39) Kozłowska, U.; Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Determination of side-chain-rotamer and virtual-bond-stretching potentials of mean force for coarse-grained simulations of protein structure and folding from AM1 energy surfaces of terminally-blocked amino-acid residues. 2. Results, comparison with statistical potentials, and implementation, *J. Comput. Chem.* **2010**, in press, DOI: 10.1002/jcc.21402.

(40) Ołdziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nanias, M.; Vila, J. A.; Khalili, M.; Arnautova, Y. A.; Jagielska, A.; Makowski, M.; Schafroth, H. D.; Kaźmierkiewicz, R.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Kang, Y. K.; Gibson, K. D.; Scheraga, H. A. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547–7552.

(41) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J. Phys. Chem. B* **2005**, *109*, 13785–13797.

(42) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model α-helical systems. *J. Phys. Chem. B* **2005**, *109*, 13798–13810.

(43) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362–2367.

(44) Kubelka, J.; Hofrichter, J.; Eaton, W. A. The protein folding 'speed limit'. *Curr. Opinion Struct. Biol.* **2004**, *14*, 76–88.

(45) Khalili, M.; Liwo, A.; Scheraga, H. A. Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. *J. Mol. Biol.* **2006**, *355*, 536–547.

(46) Cieplak, M.; Hoang, T. X.; Robbins, M. O. Thermal folding and mechanical unfolding pathways of protein secondary structures. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 104–113.

(47) Brown, S.; Fawzi, N. J.; Head-Gordon, T. Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10712–10717.

(48) Hansmann, U. H. E.; Okamoto, Y. Comparative study of multicanonical and simulated annealing algorithms in the protein folding problem. *Physica A* **1994**, *212*, 415–437.

(49) Rhee, Y. M.; Pande, V. S. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* **2003**, *84*, 775–786.

(50) Nanias, M.; Czaplewski, C.; Scheraga, H. A. Replica Exchange and Multicanonical Algorithms with the Coarse-Grained United-Residue (UNRES) Force Field. *J. Chem. Theory Comput.* **2006**, *2*, 513–528.

(51) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. Application of multiplexing replica exchange molecular dynamics method to the UNRES force field: tests with α and α + β proteins. *J. Chem. Theory Comput.* **2009**, *5*, 627–640.

(52) Nishikawa, K.; Momany, F. A.; Scheraga, H. A. Low-energy structures of two dipeptides and their relationship to bend conformations. *Macromolecules* **1974**, *7*, 797–806.

(53) Kubo, R. Generalized cumulant expansion method. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.

(54) Kolinski, A.; Skolnick, J. Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. *J. Chem. Phys.* **1992**, *97*, 9412–9426.

(55) Shen, H.; Liwo, A.; Scheraga, H. A. An improved functional form for the temperature scaling factors of the components of the mesoscopic UNRES force field for simulations of protein structure and dynamics. *J. Phys. Chem. B* **2009**, *113*, 8738–8744.

(56) He, Y.; Xiao, Y.; Liwo, A.; Scheraga, H. A. Exploring the parameter space of the coarse-grained UNRES force field by random search: Selecting a transferable medium-resolution force field. *J. Comput. Chem.* **2009**, *30*, 2127–2135.

(57) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(58) Lee, J.; Scheraga, H. A. Conformational space annealing by parallel computations: extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin. *Int. J. Quantum Chem.* **1999**, *75*, 255–265.

(59) Czaplewski, C.; Liwo, A.; Pillardy, J.; Ołdziej, S.; Scheraga, H. A. Improved Conformational Space Annealing method to treat β-structure with the UNRES force-field and to enhance scalability of parallel implementation. *Polymer* **2004**, *45*, 677–686.

(60) Gropp, W.; Lusk, E.; Skjellum, A. Parallel libraries. In *Using MPI. Portable Parallel Programming with the Message-Passing Interface*, 2nd edition; The MIT Press; Cambridge, MA, 1999; pp 157−193.

(61) Blackford, L. S.; Demmel, J.; Dongarra, J.; Duff, I.; Hammarling, S.; Henry, G.; Heroux, M.; Kaufman, L.; Lumsdaine, A.; Petitet, A.; Pozo, R.; Remington, K.; Whaley, R. C. An updated set of basic linear algebra subprograms (BLAS). *ACM Trans. Math. Soft.* **2002**, *28*−*2*, 135–151.

(62) Rakowski, F.; Grochowski, P.; Lesyng, B.; Liwo, A.; Scheraga, H. A. Implementation of a symplectic multiple-time-step molecular dynamics algorithm, based on the united-residue mesoscopic potential energy function. *J. Chem. Phys.* **2006**, *125*, 204107.

(63) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(64) Plimpton, S.; Hendrickson, B. A new parallel method for molecular dynamics simulation of macromolecular systems. *J. Comput. Chem.* **1996**, *17*, 326–337.

(65) Hendrickson, B.; Leland, R. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comput.* **1995**, *16*, 452–469.

(66) Amdahl, G. The validity of the single processor approach to achieving large-scale computing capabilities. In *Proceedings of AFIPS Spring Joint Computer Conference*, Atlantic City, NJ; AFIPS Press: 1967; pp 483−485.

CT9004068

JCTC Journal of Chemical Theory and Computation

An RNA Molecular Switch: Intrinsic Flexibility of 23S rRNA Helices 40 and 68 5′-UAA/5′-GAN Internal Loops Studied by Molecular Dynamics Methods

Kamila Réblová,*,† Zora Střelcová,‡ Petr Kulhánek,‡ Ivana Beššeová,† David H. Mathews,§ Keith Van Nostrand,§ Ilyas Yildirim,|| Douglas H. Turner,⊥ and Jiří Šponer*,†

Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 61265 Brno, Czech Republic, National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic, Department of Biochemistry & Biophysics, University of Rochester Medical Center, 601 Elmwood Avenue, Box 712, Rochester, New York 14642, Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627, and Department of Chemistry, University of Rochester, Rochester, New York 14627-0216

Received August 18, 2009

Abstract: Functional RNA molecules such as ribosomal RNAs (rRNAs) frequently contain highly conserved internal loops with a 5′-UAA/5′-GAN (UAA/GAN) consensus sequence. The UAA/GAN internal loops adopt a distinctive structure inconsistent with secondary structure predictions. The structure has a narrow major groove and forms a trans Hoogsteen/Sugar edge (*t*HS) A/G base pair followed by an unpaired stacked adenine, a trans Watson−Crick/Hoogsteen (*t*WH) U/A base pair, and finally a bulged nucleotide (N). The structure is further stabilized by a three-adenine stack and base-phosphate interaction. In the ribosome, the UAA/GAN internal loops are involved in extensive tertiary contacts, mainly as donors of A-minor interactions. Further, this sequence can adopt an alternative 2D/3D pattern stabilized by a four-adenine stack involved in a smaller number of tertiary interactions. The solution structure of an isolated UAA/GAA internal loop shows substantially rearranged base pairing with three consecutive non-Watson−Crick base pairs. Its A/U base pair adopts an incomplete cis Watson−Crick/Sugar edge (*c*WS) A/U conformation instead of the expected Watson−Crick arrangement. We performed 3.1 μs of explicit solvent molecular dynamics (MD) simulations of the X-ray and NMR UAA/GAN structures, supplemented by molecular mechanics, Poisson−Boltzmann, and surface area free energy calculations; locally enhanced sampling (LES) runs; targeted MD (TMD); and nudged elastic band (NEB) analysis. We compared parm99 and parmbsc0 force fields and net-neutralizing Na^+ versus excess salt KCl ion environments. Both force fields provide a similar description of the simulated structures, with the parmbsc0 leading to modest narrowing of the major groove. The excess salt simulations also cause a similar effect. While the NMR structure is entirely stable in simulations, the simulated X-ray structure shows considerable widening of the major groove, a loss of base−phosphate interaction, and other instabilities. The alternative X-ray geometry even undergoes a conformational transition toward the solution 2D structure. Free energy calculations confirm that the X-ray arrangement is less stable than the solution structure. LES, TMD, and NEB provide a rather consistent pathway for interconversion between the X-ray and NMR structures. In simulations, the incomplete *c*WS A/U base pair of the NMR structure is water-mediated and alternates with the canonical A−U base pair, which is not indicated by the NMR data. Completion of the full *c*WS A/U base pair is prevented by the overall internal loop arrangement. In summary, the simulations confirm that the UAA/GAN internal loop is a molecular switch RNA module that adopts its functional geometry upon specific tertiary contexts.

10.1021/ct900440t
Published on Web 01/27/2010

Introduction

RNA secondary structures comprise four basic elements such as helices, external loops (hairpin loops), internal loops, and junction loops. The first crystallographic structures of the ribosome[1−3] determined in this decade have revealed that the internal loops are structured by amazingly variable non-Watson−Crick base pairs, and many of them form recurrent structural motifs with distinct shapes.[4] (By internal loops we mean short series of nominally unpaired bases within a longer paired helix, that is, bases that do not form canonical Watson−Crick base pairs). Thus, the RNA structures can be considered as fascinating combinations of short canonical helices responsible for a major part of the thermodynamics stability and various noncanonical (non-Watson−Crick) functional elements with diverse sequences, shapes, and flexibilities. Some of the recurrent motifs are autonomous, that is, their structures are within the ribosome independent of context, while others have arrangements affected by surrounding structures, that is, exhibit induced fit binding. In the present work, we investigate one of the most salient recurrent nonautonomous RNA structural motifs that adopts its functional shape only in very specific tertiary contexts. The aim is to complement the existing structural data by analyses utilizing the available computational methods based on classical atomistic explicit solvent simulations and to establish what kind of information can in principle be gathered using modern computations for such RNA structural elements. A typical example of this element occurs in Helix 40 (H40) of the large ribosomal subunit. H40 contains a highly conserved internal loop in all three domains of life with a 5′-UAA/5′-GAN (UAA/GAN) consensus sequence (Figure 1A).[5]

This motif is present in seven internal loops of 23S rRNA and in other RNAs such as the RNase P RNAs and groups I and II introns with different degrees of conservation.[5] Despite different locations and tertiary interactions, the majority of the UAA/GAN internal loops adopt a distinctive structure with an unpaired stacked adenine, and a bulged nucleotide (N). The three conserved adenines create a characteristic cross-strand AAA stack (Figure 1A).[5] An alternative secondary structure of the loop was seen in the crystal structures of the H68 of *Escherichia coli* (*E.c.*) 23S rRNA[9] (Figure 1B) and the intact RNase P RNA from *Bacillus stearothermophilus*,[10] indicating structural plasticity of this motif.

Considering the structural data, spatial arrangement of this loop is likely to be dictated by surrounding ribosomal segments. In the ribosome, the H40 loop forms contacts with the hairpin structure between H39 and H40 by the adenines of the AAA stack via A-minor interactions.[5,11] A-minor interactions represent the most numerous and highly conserved tertiary interactions in large structured RNAs and ribonucleoproteins. The bulged nucleotide is involved in tertiary contacts as well. The bacterial H40 internal loop is additionally a part of the binding site of the ribosomal protein L20.[12] Recently, the boxed motif in Figure 1A was classified as a UA_handle submotif, which is a highly versatile nonautonomous common RNA building block.[8]

The structure of the UAA/GAA motif flanked by Watson−Crick base pairs was determined in solution by NMR spectroscopy.[13] There are striking differences between the X-ray ribosomal H40 and the solution structure (Figure 1A,C). The pairing is restructured so that there are no unpaired or bulged bases. The solution structure is much more consistent with the arrangement expected on the basis of 2D thermodynamics prediction, except that the standard 2D predictions would propose a canonical A−U base pair instead of the observed A/U base pair. The solution structure has a considerably wider (more open) major groove compared to the X-ray H40 UAA/GAA segment, which has a narrow (closed) major groove and wide minor groove (Figure 1). However, this feature may be a trivial consequence of the NMR structure refinement utilizing force field calculations. Thus, the functional structure seen in the ribosomes differs from the presumably intrinsically preferred arrangement seen in solution and is likely stabilized by tertiary and quaternary interactions.[13]

X-ray and NMR studies of RNAs can be complemented by molecular dynamics (MD), which provides dynamic data and additional insight into the structure.[14−30] For instance, MD simulations can characterize isolated RNA building blocks independent of their structural context. The simulations capture intrinsic stochastic fluctuations of geometries and reveal intrinsic elastic properties that are important for function.[31−34] In addition, simulations can disclose whether an RNA building block is entirely internally relaxed or is deformed due to its interactions with the surroundings. In the latter case, simulations can reveal rapid structural relaxation toward low-energy conformations.[35,36] These insights are unique, and they would be hard to obtain by other methods. Thus, despite being limited by force field approximations and time scale, MD simulations provide valuable data that can complement experiments.[37−42]

In the present study, we run explicit solvent MD simulations on isolated internal loops of H40 starting with solution structure determined by NMR and the structure in ribosomes as determined by X-ray diffraction (Figure 1A,C). The study is complemented by MD simulation of the X-ray loop of H68 from the ribosome (Figure 1B). The main aim was to characterize base pairing and local arrangement of the loop on the nanosecond time scale to better understand its structural plasticity. The simulations were supplemented by free energy calculations that extract free energy directly from the trajectories. We also utilized locally enhanced sampling (LES), nudged elastic band (NEB), and targeted MD (TMD) to investigate the pathway for the conformational change between the NMR and X-ray structures. Standard net-neutralizing Na^+ simulations with the parm99 AMBER force field[43] were compared with the parmbsc0[44] force field and

* Corresponding author tel.: +420 541517250 (K.R.), +420 541517133 (J.S.); fax: +420 541212179 (K.R.), +420 541212179 (J.S.); e-mail: kristina@physics.muni.cz (K.R.), sponer@ncbr.chemi.muni.cz (J.S.).

† Academy of Sciences of the Czech Republic.

‡ Masaryk University.

§ University of Rochester Medical Center.

|| Department of Physics and Astronomy, University of Rochester.

⊥ Department of Chemistry, University of Rochester.

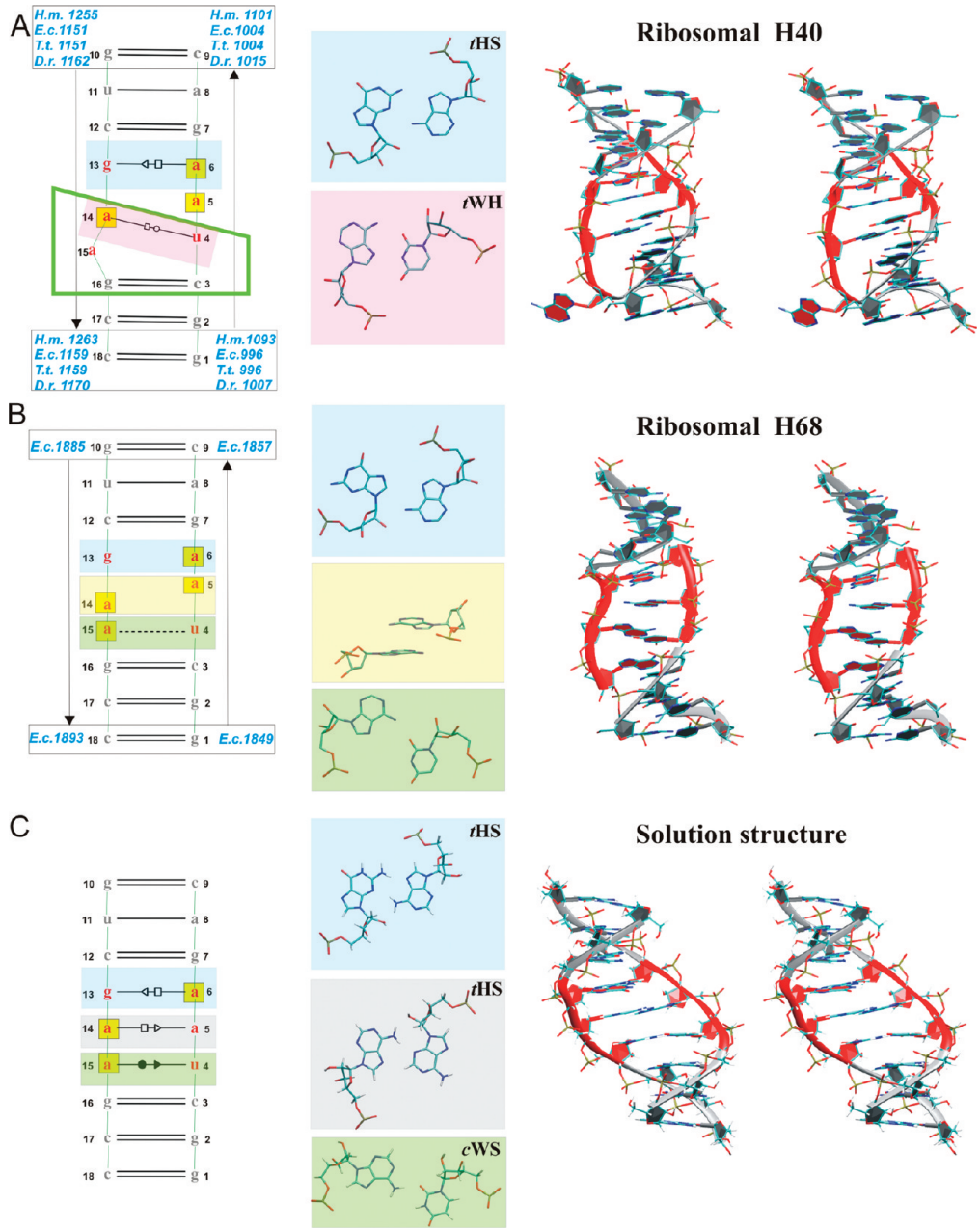


Figure 1. The 2D structures (left) and 3D stereo views (right) of studied segments including non-Watson−Crick base pairs in the internal UAA/GAA loop (middle). (A, left) The 2D X-ray structure of H40 with unified sequence flanking the internal loop (see the Materials and Methods). (A, middle) Sheared A/G and rH U/A base pairs. (A, right) Stereo view of *E. coli* H40 X-ray structure. (B, left) The 2D X-ray structure of *E. coli* 23S rRNA H68 exhibiting an alternative conformation of the UAA/GAA motif with a unified canonical flanking sequence. The black dashed line indicates a single H bond. (B, middle) Unpaired G and A bases, stacking middle adenines, and single-bonded A/U base pair. (B, right) Stereo view of this structure. (C, left) The 2D NMR structure. (C, middle) Sheared A/G, sheared A/A, and incomplete *c*WS A/U base pairs. (C, right) Stereo view of the NMR structure. In all parts, bases of the UAA/GAA internal loop are in red, 3D structures are colored accordingly, hydrogens are not shown in the X-ray structures, bases in yellow boxes in the 2D structures are involved in stacking, and the marks between the bases indicate the base paring family according to the Leontis and Westhof classification (*t*HS = trans Hoogsteen/Sugar edge A/G or A/A, known also as "sheared" base pairs; *t*WH = trans Watson−Crick/Hoogsteen U/A, known also as reverse Hoogsteen (rH) base pair; and *c*WS = cis Watson−Crick/Sugar edge A/U).[6,7] X-ray nucleotide numbers are in blue; NMR numbers are in black. The green rectangular trapezium for H40 structure marks bases forming the UA_handle submotif.[8]

excess salt KCl simulations. This study had two purposes: first, to investigate an important RNA modular block stabilized by tertiary contacts that appears to act as a flexible RNA structural switch; second, to test the capability of explicit solvent simulations and some auxiliary techniques to describe structural dynamics of RNA. The combination of methods provides interesting qualitative insights into the intricate properties of the UAA/GAN RNA internal loop.

Our standard simulations show relaxation of the ribosomal H40 loop. The X-ray conformation opens significantly and adopts an arrangement that resembles the solution structure. However, the X-ray secondary structure of the loop was

maintained on the 200−300 ns time scale despite numerous local disturbances evidenced by the simulations. Thus, spontaneous transition of the H40 loop to the solution structure (presumably the global free energy minimum conformation) was not achieved. Disruption of the H40 X-ray secondary structure was detected in the LES simulations, where the NMR-like secondary structure was sampled, albeit not dominantly. In contrast, almost perfect transition was observed in the standard simulation of the H68 loop, which is probably not as deformed as the ribosomal H40 loop due to the reduced number of tertiary contacts. The free energy calculations clearly show that the solution structure of the UAA/GAA internal loop is more stable compared to its ribosomal geometries and reveal a free energy change of the conformational transition between the H68 and NMR structures. Finally, LES, NEB, and TMD calculations allowed us to propose a plausible mechanism by which the solution conformation could rearrange into the "ribosomal" H40 X-ray geometry. Both parmbsc0 and KCl simulations are qualitatively in agreement with standard net-neutralizing Na^+ parm99 simulations. The major groove opening is, however, reduced compared to that of the Na^+ parm99 simulations.

Materials and Methods

Starting Structures. The H40 X-ray structure was taken from the available 50S subunits, that is, from the archaeal subunit of *Haloarcula marismortui* (*H.m.*; pdb code 1S72)[45] and from bacterial subunits of *E.c.* (2AW4),[9] *Deinoccocus radiodurans* (*D.r.*; 1NKW),[2] and *Thermus thermophilus* (*T.t.*; 2J01)[46] (Figure 1A and Figure S1, Supporting Information). We have compared conformations of 23S rRNA H40 from the recent structures of Ramakrishnan that include EF-G[47] and EF-Tu[48] to the *T.t.* X-ray structure used in the present study. No visible structural differences were detected among the structures. The RMSDs between the X-ray structure used in the present study and the newly released structures were below 1 Å, so they are identical. The H40 NMR structure was taken from pdb 2H49 (we utilized model 1;[13] Figure 1C). Both X-ray and NMR structures have a UAA/GAA internal loop (only the *H.m.* helix contains the UAA/GAG sequence) flanked by various Watson−Crick base pairs (Figure 1). In our study, the Watson−Crick base pairs in the X-ray structures (Figure S2, Supporting Information) were mutated to match the NMR structure (Figure 1C), with an additional UAA/GAG → UAA/GAA substitution introduced for the *H.m.* structure. Thus, all systems have the same sequence, which allowed us to compare free energies in the studied systems (see below). Although all simulated structures have identical sequences, the starting structures still reflect some local structural differences of the X-ray structures. We used the NMR nucleotide numbering (1−18) for X-ray H40s throughout the text to unify the description of the systems (Figure 1). Original X-ray numbers are indicated in Figures 1 and S2 (Supporting Information) and Table 1.

All the X-ray H40s have the same secondary structure and almost identical local geometry (Figures 1 and S1 and S2, Supporting Information).

The X-ray internal loop comprises a sheared A6/G13 pair and rH U4/A14 pair, a bulging A15 base, and an A5 base stacked within the stem without a base pairing partner on the other strand (Figure 1A). A5 forms an intrastrand stack with the A6 base and an interstrand "cross"-strand stack with the A14 base, resulting in the characteristic AAA stack (Figure 1A). These interactions shape the internal loop into a specific arrangement exhibiting a broadened minor groove and narrow major groove (∼6−8 Å) stabilized by a base−phosphate (BPh) interaction.[49] The *E.c.* X-ray structure of H40 exhibits a bifurcated binding mode (peculiar alternative of base phosphate interaction type 4BPh) in which N2 and N1 of G13 bind to the same anionic oxygen of the phosphate group.[49] In particular, there are G13(N1)−A5(O2P) and G13(N2)−A5(O2P) H bonds (Figure 2). In contrast, *H.m.*, *D.r.*, and *T.t.* X-ray structures exhibit base phosphate interaction type 5BPh, including only the G13(N1)−A5(O2P) H bond (Figure 2).[49]

The NMR internal loop also shows the sheared A6/G13 base pair seen in the H40 X-ray structures; however, the loop contains two additional single hydrogen bonded noncanonical base pairs: a sheared A14/A5 and *c*WS A15/U4 base pair (Figure 1C). The latter base pair is classified as a *c*WS,[6] although it contains only one direct A(N6)−U(O2) H bond. The simulations reveal that the other interaction characteristic for *c*WS A/U base pairs, the A(N1)−U(O2′) interaction, is in fact water-mediated. The major groove of the NMR structure is wide (open, ∼17 Å; Figure 1C), and it does not have any base−phosphate contacts across the groove.[13] The NMR structure reveals another AAA stack (Figure 1C) with a cross-strand A6/A14 and intrastrand A14/15 interactions.

The X-ray UAA/GAA loop of H68 was taken from *E.c.* 50S (2AW4)[9] (Figure 1B). Base pairs above and below the loop were mutated according to the Watson−Crick base pairs of the NMR structure (Figure 1). The internal loop of H68 consists of unpaired G13 and A6 bases (Figure 1B) where the G13 base is in the syn conformation. In the ribosomal H40 and in the solution structure, the G13 is in anti orientation, and in both of these structures it forms a sheared A6/G13 pair (see Figure 1). Furthermore, the loop of H68 comprises stacked middle adenines A5 and A14 and an imperfect *c*WS A15/U4 base pair stabilized by a single H bond (A(N6)−U(O2); Figure 1B). In addition, the A5 and A14 bases form intrastrand stacks with adjacent adenines A6 and A15, respectively, resulting in a four-adenine stack (Figure 1B). The major groove is a little wider (∼9−10 Å) compared with the H40 geometry.

Ribosomal Contacts of H40. The studied UAA/GAN H40 internal loops are involved in identical A-minor interactions in all four 50S subunits. The three adenines of the cross-strand AAA stack interact with two consecutive highly conserved C═G base pairs of the hairpin between H39 and H40.[5] In particular, the adenine of the sheared A/G pair forms a type I A-minor interaction with one C═G pair, while the A5 adenine forms a type II A-minor interaction with the next C═G pair (G═C in *T.t.*). Finally, the adenine of the rH U/A base pair forms a tilted variant of A-minor type I also with the second G═C base pair (Figure 3).[5]

In bacterial ribosomes, the minor groove side of the UAA/GAN motif interacts with the hairpin between H39 and H40, while the major groove side interacts with ribosomal protein

Table 1. Survey of Performed Simulations[a]

organism	simulated segment (original experimental numbering)	simulation name	resolution (Å) and pdb code	length of simulation (ns)	RMSD[b] (Å)	force field
		Standard MD Simulations with Net-Neutralizing Na^+ Atmosphere				
E.c.	996−1004, 1151−1159	MD_Ec_99	3.5, 2AW4	350[c]	4.4 ± 1.2	Parm99
E.c.	996−1004, 1151−1159	MD_Ec_bsc0	3.5, 2AW4	100	2.3 ± 0.4	Bsc0
H.m.	1093−1101 1255−1263	MD_Hm_99	2.4, 1S72	250[d]	3.5 ± 1.0	Parm99
H.m.	1093−1101, 1255−1263	MD_Hm_bsc0	2.4, 1S72	100	2.7 ± 0.5	Bsc0
D.r.	1007−1015, 1162−1170	MD_Dr_99	3.1, 1NKW	200	4.8 ± 0.7	Parm99
D.r.	1007−1015, 1162−1170	MD_Dr_bsc0	3.1, 1NKW	100	4.3 ± 0.8	Bsc0
T.t.	996−1004, 1151−1159	MD_Tt_99	2.8, 2J01	200[d]	2.4 ± 0.7	Parm99
T.t.	996−1004, 1151−1159	MD_Tt_bsc0	2.8, 2J01	100	2.3 ± 0.5	Bsc0
E.c.	996−1004, 1151−1159	MD_A5U[e]	3.5, 2AW4	100	2.4 ± 0.5	Parm99
E.c.	996−1004, 1151−1159	MD_A14U[f]	3.5, 2AW4	100	2.2 ± 0.3	Parm99
H.m.	1093−1101, 1255−1263	MD_A14G_U4C[g]	2.4, 1S72	50	1.9 ± 0.4	Parm99
E.c.	996−1004, 1151−1159	MD_nosalt[h]	3.5, 2AW4	150	4.3 ± 1.0	Parm99
E.c.	996−1004, 1151−1159	MD_400K[i]	3.5, 2AW4	20	6.7 ± 3.2	Parm99
E.c.	996−1004, 1151−1159	MD_400K[j]	3.5, 2AW4	20	5.1 ± 2.7	Parm99
E.c.	996−1004, 1151−1159	MD_LES_Ec[k]	3.5, 2AW4	80	4.2 ± 1.7	Parm99
N/A	1−18	MD_NMR_99	N/A, 2H49	200	1.6 ± 0.3	Parm99
N/A	1−18	MD_NMR_bsc0	N/A, 2H49	100	1.2 ± 0.2	Bsc0
N/A	1−18	MD_NMR_restr[l]	N/A, 2H49	200	1.7 ± 0.3	Parm99
E.c.	1885−1893, 1849−1857	MD_H68	3.5, 2AW4	100	3.9 ± 0.5	Parm99
		Standard MD Simulations in Excess of KCl				
E.c.	996−1004, 1151−1159	MD_Ec_K1[m]	3.5, 2AW4	100	2.1 ± 0.5	Parm99
E.c.	996−1004, 1151−1159	MD_Ec_K2[m]	3.5, 2AW4	100	2.0 ± 0.3	Bsc0
E.c.	996−1004, 1151−1159	MD_Ec_K3[n]	3.5, 2AW4	100	2.9 ± 0.4	Parm99
N/A	1−18	MD_NMR_K[m]	N/A, 2H49	100	1.5 ± 0.3	Parm99
		LES Simulations				
E.c.	996−1004, 1151−1159	LES_Ec	3.5, 2AW4	60	6.7 ± 2.2	Parm99
D.r.	1007−1015, 1162−1170	LES_Dr	3.1, 1NKW	40	5.9 ± 1.3	Parm99
T.t.	996−1004, 1151−1159	LES_Tt	2.8, 2J01	40	6.4 ± 1.3	Parm99
H.m.	1093−1101, 1255−1263	LES_Hm	2.4, 1S72	40	6.4 ± 2.0	Parm99

[a] The sequence of the simulated molecules was unified to match the sequence used in the NMR study (see the Materials and Methods and Supporting Information). [b] RMSD values are calculated along the trajectory for the individual snapshots with respect to the starting structure. [c] Due to disruption of the structure, we considered only the 0−300 ns trajectory portion in the analyses. [d] Due to disruption of the structure, we considered only 0−150 ns in the analyses. [e] Simulation run with A5U mutation. [f] Simulation run with A14U mutation. [g] Simulation run with A14G and U4C mutations. [h] Simulation run under no-salt conditions. [i] Simulation run at 400 K (NVT). [j] Simulation run at 400 K (NPT). [k] Standard MD simulation that started from the NMR-like conformation observed in the LES_Ec simulation. [l] Simulation run with restraint, which enforced a direct A(N1)−U(O2′) H bond of the *c*WS A/U base pair (instead of the water-mediated one) for 10 ns. [m] Simulation run with Dang's parameters for K^+ and Cl^- (see the Materials and Methods). [n] Simulation run with Joung and Cheatham's parameters for K^+ and Cl^- (see the Materials and Methods).

L20 (Figure 3). In contrast, in the archaeal *H.m.* 50S, the ribosomal protein L20 is substituted by ribosomal protein L30, which interacts with the minor groove side of the UAA/GAN internal loop (Figure 3). Moreover, the *H.m.* UAA/GAN has contact with H25 (Figure 3). Apparently, the UAA/GAN internal loop of H40 exhibits different interactions on its minor groove side in bacteria and archaea.

As noted by Lee et al.,[5] the UAA/GAN motifs in other helices adopting the same 2D/3D arrangement are also involved in extensive molecular contacts resembling those of H40 UAA/GAN. The alternative 2D/3D UAA/GAN conformation (H68 of *E.c.* 23S rRNA and in intact RNase P RNA) contacts just one RNA helix (Figure 3). This is another indication that structural context may alter the geometry of the UAA/GAN motif.

Standard MD Simulations. Standard explicit solvent simulations were carried out (at 300 K) using the pmemd module of AMBER 9.0[50,51] and force field parm99[43] version of the Cornell et al. force field[52] on a time scale of 200+ ns each. Control simulations of 100 ns were run with parmbsc0.[44] Parmbsc0 is the latest reparametrization of the Cornell et al. force field aimed to stabilize B-DNA simulations by penalizing substates with γ-trans backbone topologies. While

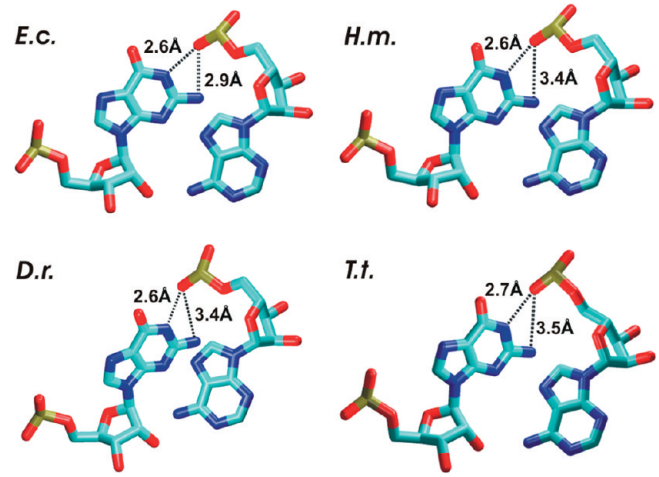


Figure 2. Base phosphate interactions observed in the ribosomal X-ray structures of H40. The *E.c.* structure exhibits a bifurcated binding mode (base phosphate interaction type 4BPh) in which N2 and N1 of G13 bind to the same anionic oxygen of the phosphate group A5(O2P). The *H.m.*, *D.r.*, and *T.t.* structures exhibit only the G13(N1)−A5(O2P) H bond, which represents base phosphate interaction type 5BPh. The differences might reflect limits of the resolution of the experimental structures.

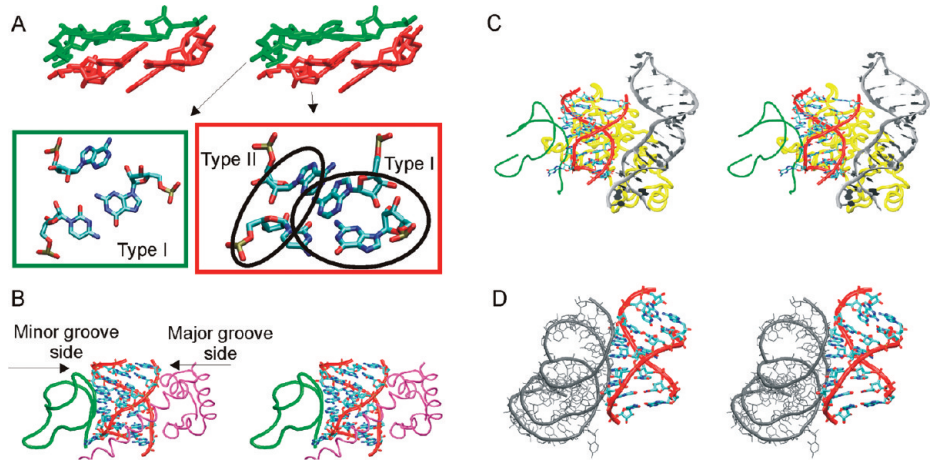


Figure 3. (A) Stereo view of three adenines of the UAA/GAA motif from *E.c.* H40 forming an AAA stack which interacts with two C═G base pairs from the hairpin between H39 and H40 via A-minor interactions. The C═G pair and the corresponding interacting adenine(s) are highlighted with the same color. Details of these interactions are visualized below the stereo view in corresponding green and red boxes, including a description of the A-minor interaction type. (B) Stereo view of bacterial *E.c.* H40 interacting with the hairpin between H39 (in green) and H40 and ribosomal protein L20 (in magenta). (C) Stereo view of archaeal *H.m.* H40 interacting with the hairpin between H39 (in green) and H40, ribosomal protein L30 (in yellow), and H25 (in gray). (D) Stereo view of bacterial *E.c.* H68 (exhibiting the alternative conformation of UAA/GAN internal loop) interacting with H75 (in gray).

parmbsc0 achieves a considerably improved performance at predicting DNA backbone conformations as compared to parm99, no systematic comparative testing was done for RNA with both parm99 and parmbsc0 until now. This work supports the viability of both force fields for RNA simulations. An overview of all simulations is given in Table 1. The total length of performed simulations was 3.1 μs. The simulated molecules were neutralized by standard AMBER Na^+ monovalent cations (radius 1.868 Å and well depth 0.00277 kcal/mol),[53] initially placed using the xleap module of AMBER at the most negative sites around the solute. RNA molecules were solvated by a TIP3P water box to a depth of 10 Å. Net-neutralization corresponds to a cation concentration of ~0.15−0.2 M. Some control simulations were carried out in KCl with ~0.2 M excess salt concentration (Table 1). For this purpose, we used either modified parameters for K^+ (radius 1.705 Å and well depth 0.1936829 kcal/mol) and Cl^- (radius 2.513 Å and well depth 0.0355910 kcal/mol), which prevents salt crystallization at low to medium salt concentrations,[54] or Dang's parameters for K^+ (radius 1.87 Å and well depth 0.1 kcal/mol)[55] and Cl^- (radius 2.47 Å and well depth 0.1 kcal/mol)[56] together with an SPC/E water box.[57,58] The simulations were carried out using the particle mesh Ewald (PME) technique with a 9 Å nonbonded cutoff and a 2 fs integration time step. The trajectory was saved once a picosecond. Equilibration and production phases were carried using standard protocols.[20] Trajectories were analyzed using the ptraj module of AMBER, and structures were visualized using the VMD molecular visualization program, http://www.ks.uiuc.edu/Research/vmd/ (accessed January 2010).[59] The figures were prepared using VMD. The stability of stacking interactions within the AAA stack of H40 was monitored by calculating the van der Waals interaction energies between the A5 and A14, A5 and A6, and A5 and G13 bases utilizing the anal module of AMBER.

To enhance sampling of H40, two simulations were carried out at an elevated temperature (400 K; Table 1). The system was gradually heated from 300 to 400 K during the first 100 ps using NPT conditions (constant pressure ensemble). The production runs were continued at 400 K using both NPT and NVT (constant volume ensemble). There are no clear guidelines whether NVT or NPT simulations should be preferred, and in fact, both approaches have drawbacks.[60] Elevated temperature simulations were previously successfully applied to studies of a base substitution in an RNA frameshifting pseudoknot.[61] Additionally, a "no-salt" simulation of H40 in which cations were omitted from the simulation box was carried out (Table 1). The missing counterions were substituted by a net-neutralizing plasma representing a uniform neutralizing charge distribution over the box. This feature is implemented in the AMBER program package for use with the PME method and guarantees the neutrality of the system.[62] The aim of the no-salt simulation was to destabilize the simulated structure.

Locally Enhanced Sampling (LES) Simulations. LES simulations[63−65] were carried out using the sander program of AMBER 9.0[50,51] to enlarge sampling of the internal loop started from the X-ray structures of H40 UAA/GAA internal loops. The addles module of AMBER was used to split the internal loop region into five independent copies; that is, residues 4−6 and 13−15 were copied five times. Force field parameters for the copies were scaled, which lowered the energy barriers on the potential energy surface and increased the flexibility of the given region. The equilibration and production phases were carried out using standard protocols.[20] Heating during the equilibration phase was continued up to 300 K. The LES method along with explicit solvent simulations were successfully utilized in studies of several nucleic acids systems.[20,66−68]

Free Energy Calculations. The molecular mechanics, Poisson−Boltzmann, and surface area method (MM-

PBSA)[69,70] implemented in AMBER 9.0[50,51] was used for free energy analysis of the explicit solvent MD trajectories. This method is based on a continuum solvent approach that replaces the explicit solvent and utilizes snapshots directly from the simulation. Here, it was employed to estimate the total free energy of H40 along MD_Ec_99, MD_Hm_99, MD_Dr_99, and MD_Tt_99 simulations and of H68 along the MD_H68 simulation (see Table 1). Total free energy of the solution structure along the MD_NMR_99 simulation was also obtained. The calculations of MM-PBSA energy included calculations of molecular mechanics energy (EMM) and solvation energy (EPBSA). EMM was calculated by the sander module of AMBER (explicit solvent and ions were not included) with parm99.[43] EPBSA is composed of two types of contributions, electrostatic (EPB) and nonpolar (ESA). The electrostatic part was calculated with a numerical solver with the Poisson−Boltzmann method implemented in the PBSA program.[71] The nonpolar part depends on solvent-accessible surface area, which was calculated by the molsurf program implemented in AMBER.[72] Conformational entropy was obtained using the nmode module of AMBER 9.0,[50,51] which performs normal-mode analysis[73] to predict the conformational entropy. The program provides a total solute entropic term as a sum of translational, rotational, and vibrational entropic contributions. All free energy terms were derived using each consecutive 20th snapshot.

Nudged Elastic Band (NEB) Method. The NEB method was employed to investigate the conformational change pathway for the transformation between the NMR and X-ray structures. The original NMR and X-ray structures were energy-minimized using standard methods with AMBER 10.0[74] and parm99.[43] The potential energy for the minimized NMR structure was −4087.4 kcal/mol, and for the X-ray structure was −4000.1 kcal/mol. These structures were used as end points in NEB calculations.[75−77] The initial NEB pathway consisted of 16 NMR structures followed by 16 X-ray structures. Twenty-one NEB trajectories were calculated using a simulated annealing protocol and varying the random number seed. The simulated annealing protocol (Table S1, Supporting Information)[78] involved quickly heating the system to 1000 K, followed by slow cooling and finally quenched dynamics to remove any remaining kinetic energy from the system.

Targeted MD (TMD). TMD was used as implemented in AMBER 9.0[50,51] to perform a forced conformational transition between the NMR and X-ray structures. The NMR structure was equilibrated using the standard protocol,[20] and then it was used as a starting point (reactant structure) and the equilibrated *E.c.* X-ray structure of H40 as the end point (target structure). The reaction coordinate was defined as the RMSD of the internal loop (residues: 4−6, 13−15) between the instantaneous reactant structure and the fixed target (product) structure. Since the RMSD is dependent on a group of atoms, which are used for the best fit to the target structure, we ran two TMD simulations with different initial settings. In the MD_TMD_1 simulation, modest positional restraints (0.01 kcal·mol^{-1}·Å^{-2}) were applied to terminal WC base pairs (residues: 1, 2, 8−11, 17, 18), which were simultaneously used for the best fit. Simulation was carried out in 20 1-ns-long windows. In each window, the molecule was forced to target RMSD, which gradually decreased (in each window about ∼0.4 Å increment; Figure S3, Supporting Information). In the MD_TMD_2 simulation, the positional restraints were not used, and the best fit was carried out over the whole structure (i.e., over residues 1−18). This simulation included only 10 1-ns windows where target RMSD gradually decreased by ∼0.4 Å increments (Figure S3, Supporting Information). Both simulations were run in explicit solvent at 300 K using NPT conditions. Control simulations for both MD_TMD_1 and MD_TMD_2 were run with different random number seeds. The force constant was set to 0.1 kcal·mol^{-1}·Å^{-2} in both simulations. Apart from tracking the conformational transition, we employed the weighted histogram analysis method[79] (version 1.0) to estimate the free energy profile of the conversion. Note that targeted MD is substantially affected by the imposed path[80] such that any large-scale conformational changes like those in refs 22 and 81 as well as in the present study should always be reviewed carefully and only be viewed as crude estimates of the real transitions.

Results

Standard MD Simulations. *Geometry of the Ribosomal H40 Relaxes in parm99 MD Simulations.* Definitely the most striking feature in the 350 ns MD_Ec_99 simulation was considerable opening of the structure due to widening of the major groove (Figures 1A, right, and 4A).

In the course of the simulation, the major groove width was monitored by two interphosphate distances (11P−4P and 12P−3P; Figure 4A). In the time period from 0 to 30 ns, the major groove width oscillated around 9 Å (Figure 4B); then it rapidly increased up to ∼16 Å and oscillated around this value until ∼100 ns (Figure 4B). In the 100−300 ns time period, the major groove width fluctuated around 20 Å (Figure 4B). The opening, which was also seen in other simulations of X-ray H40, was coupled with the disruption of the BPh interaction.[49] The two noncanonical base pairs of the internal loop showed instabilities during the simulation. Particularly, opening events of both H bonds of the sheared A/G pair were detected (Table 2).

Disruption of this pair was seen in the 133−297 ns time period, during which the A6 base flipped out of the helix. Changes were also detected for the rH U/A base pair. An opening event and eventual disruption was seen for the U(N3)−A(N7) H bond (Table 2). The cross-strand A5A14 stack exhibited fluctuations in the 0−100 ns time period; however, in the rest of the simulation, it was essentially stable (Figure 5).

Larger changes were found for the A5A6 intrastrand stack. In the 0−100 ns time period, the A5 base alternatively stacked between A6 and G13 and, afterward, established stable stacking with G13 (Figure 5). The A15 bulge fluctuated outside the helix over the whole simulation. Its insertion into the stem was obstructed by A14, which was involved in cross-strand stacking. The MD_Ec_99 simulation was extended up to 350 ns. However, the canonical segment (residues 7−12) including the sheared A6/G13 base pair was disrupted at ∼300 ns (Figure S4, Supporting Information).

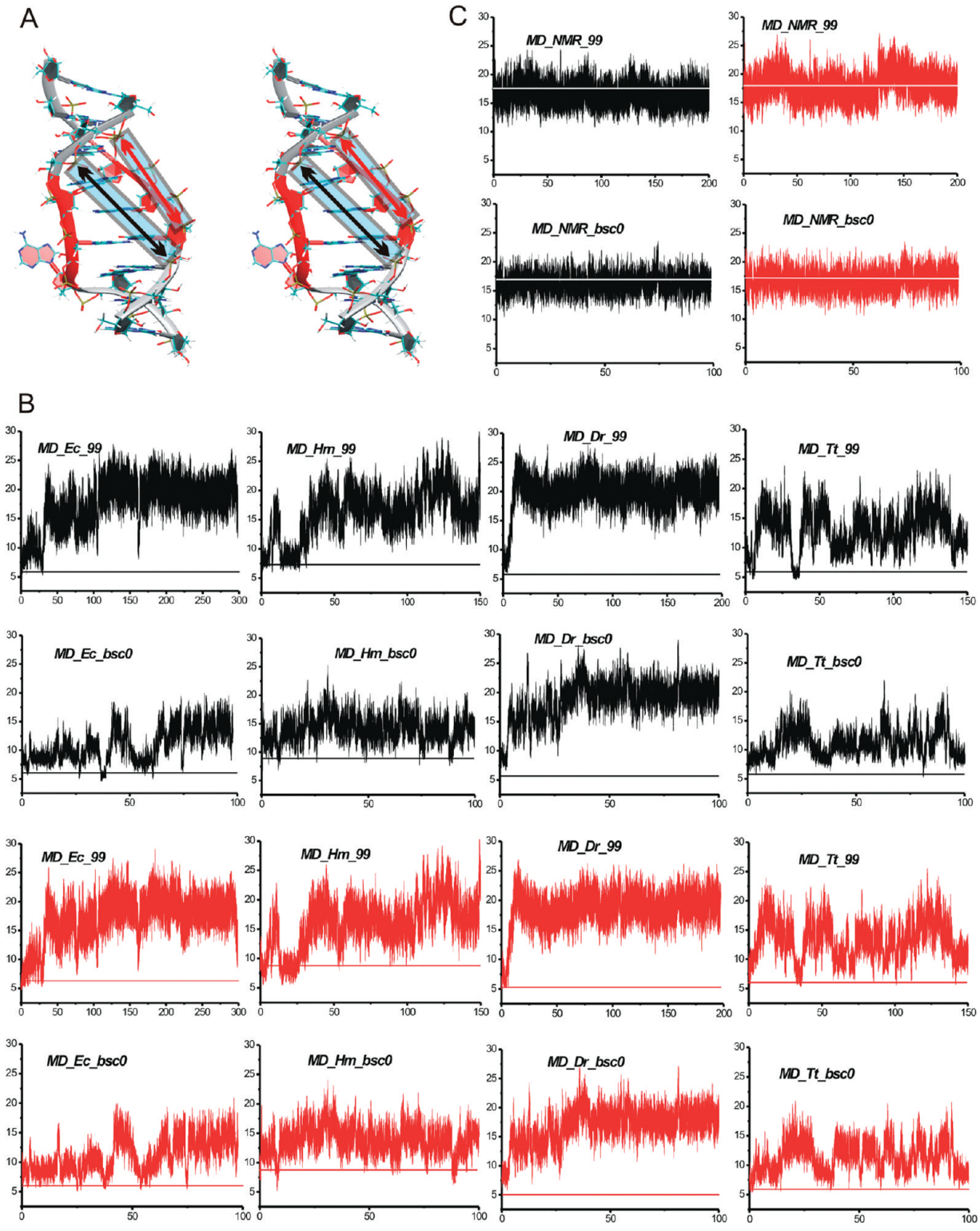


Figure 4. (A) Stereo view of the averaged 55−57 ns MD structure of the simulated X-ray H40 UAA/GAA internal loop from the MD_Ec_99 simulation. The structure exhibits a wider (more open) major groove compared to the original geometry (see Figure 1A). Monitored interphosphate distances across the major groove are indicated by black (12P-3P) and red (11P-4P) arrows in blue transparent boxes. (B) Time courses of two interphosphate distances (12P−3P in black and 11P−4P in red) in standard MD simulations of X-ray H40 run with the parm99 force field and in control simulations run with the parmbsc0 force field, all with net-neutralizing Na^+ (Table 1). The *x* axis stands for time (in nanoseconds), while the *y* axis stands for interphosphate distance (in angstroms). Horizontal lines show experimental distances. (C) Time course of two interphosphate distances (12P−3P in black and 11P−4P in red) in a standard MD simulation of the NMR structure run with the parm99 force field and in a control simulation run with the parmbsc0 force field.

Thus, the simulated structure was ultimately lost, and the 300−350 ns time period was not considered in our analyses.

The MD_Hm_99 and MD_Dr_99 simulations showed a picture similar to the MD_Ec_99 simulation, that is, a marked

Table 2. Base Pairing Changes Detected in the Standard Simulations Performed with the X-Ray H40 UAA/GAA Structure[a]

	sheared A/G pair		reverse Hoogsteen U/A pair	
simulation name	A(N7)−G(N2), H bond	A(N6)−G(N3), H bond	U(O2)−A(N6), H bond	U(N3)−A(N7), H bond
MD_Ec_99	75−78 ns[oe] 133−297 ns[d]	33−40 ns[oe] 65−78 ns[oe] 133−297 ns[d]	stable	89.2−90.1 ns[oe] at 101[d]
MD_Ec_bsc0	f	stable	stable	18.5−23.2 ns[oe] 41.3−46.8 ns[oe] at 65 ns[d]
MD_Hm_99	at 150 ns[d]	39−45 ns[oe] 45−97[f] 97−109[oe] 150 ns[d]	stable	56−81 ns[oe]
MD_Hm_bsc0	1−5 ns[oe] 90−94 ns[oe]	f	stable	5−15 ns[oe] 26−37 ns[oe] 75−96 ns[oe]
MD_Dr_99	stable	f	stable	at 7 ns[d]
MD_Dr_bsc0	stable	f	stable	at 30 ns[d]
MD_Tt_99	63−65 ns[oe]	63−65 ns[oe]	stable	70 − 122 ns[oe]
MD_Tt_bsc0	58−69 ns[oe]	58−69 ns[oe]	stable	13−17 ns ns[oe] 24−32 ns[oe] 60−78 ns[oe] 83−96 ns[oe]
MD_Ec_K1	f	stable	77−80 ns[oe]	4−22 ns[oe] 77−80 ns[oe] 96−98 ns[oe]
MD_Ec_K2	stable	f	stable	11−28 ns[oe]
MD_Ec_K3	stable	stable	stable	8−9 ns[oe] 16−20 ns[oe]

[a] oe, d, and f stand for temporary opening, disruption (until the end of the simulation), and considerable fluctuations, respectively.

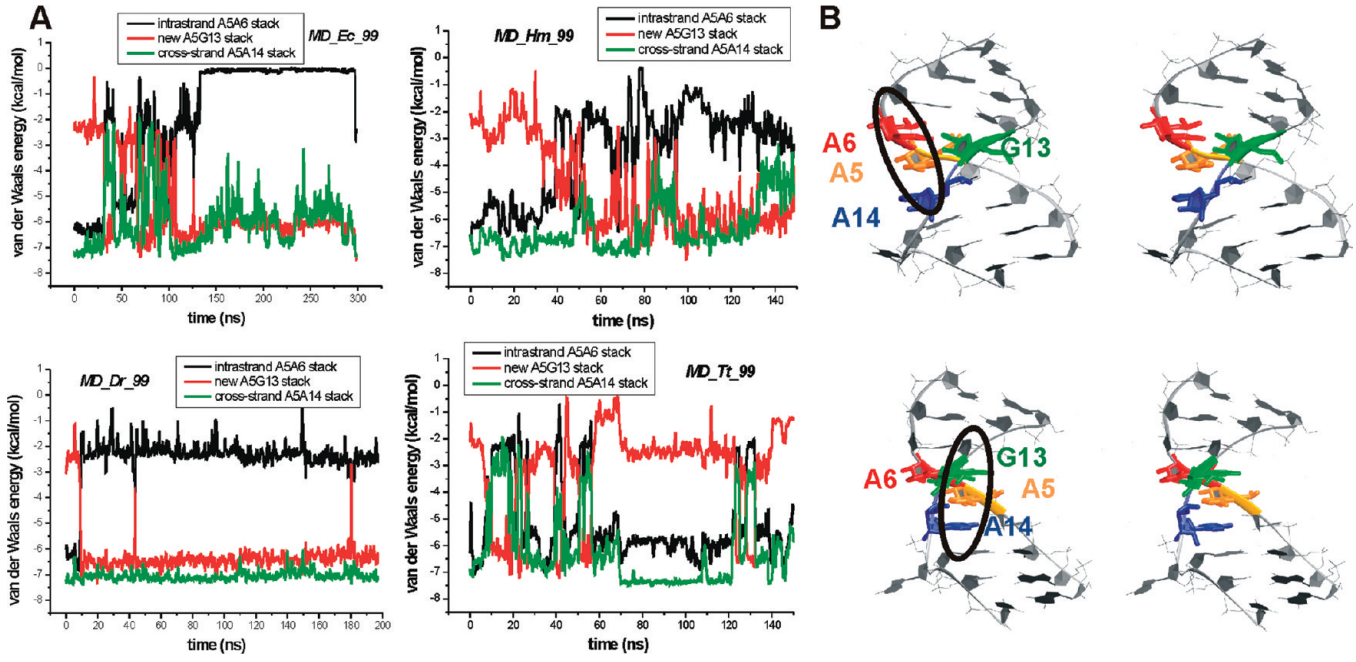


Figure 5. (A) Time courses of the van der Waals interaction energy calculated between bases forming A5A6, A5A14, and A5G13 stacks in the standard simulations of an X-ray H40 run with the parm99 force field. (B) Stereo views of the X-ray H40 UAA/GAA internal loop with colored bases forming stacks. (Top) The original stacking pattern highlighted by the black oval; that is, A5 forms an intrastrand stack with A6 and, simultaneously, a cross-strand stack with A14. (Bottom) The stacking pattern formed in the course of the simulations (highlighted by the black oval) where A5 stacks with G13 and also with A14.

opening of the major groove (Figure 4B), fluctuation of the A15 base outside the helix, instability of the two non-Watson−Crick base pairs (Table 2), and changes in base stacking (Figure 5). Hence, full description of these simulations is given in the Supporting Information. In the MD_Tt_99 simulation, the major groove did not convert into a permanently open conformation in contrast to the previous simulations. The major groove width oscillated, and it was stabilized by a temporarily formed X-ray G13(N1)−A5(O2P) H bond (Figure 2). Apart from this contact, an additional H bond formed between G13(N2) and A5(O2P), which however oscillated in a larger range compared to the G13(N1)−A5(O2P). This binding mode in which N2 and N1 of G bind to the same anionic oxygen of the phosphate group represents

base phosphate interaction type 4BPh, which can be seen in the *E.c.* X-ray structure of H40 (Figure 2). In the course of the MD_Tt_99 simulation, the original 5BPh interaction thus alternated with the 4BPh interaction. Both the intrastrand and cross-strand stack fluctuated, but replacement of the A5A6 stack by the A5G13 stack was not seen. Similarly to the other simulations, instabilities of the two noncanonical base pairs were detected (Table 2). Details of this simulation are also indicated in the Supporting Information.

MD Simulations of the Ribosomal H40 with Mutated Residues. The A5A14 cross-strand stack is likely key for maintaining the stability of the X-ray loop. Thus, we carried out the MD_A5U simulation with an A5U mutation aimed at destabilizing this stack (Table 1). The simulation showed expulsion of the uracil out of the stem and subsequent stacking of the sheared pair and the rH pair. This caused shortening of the helix by one base pair level and formation of a more compact and presumably more stable structure. This substitution apparently destabilizes the functional structure of the UAA/GAN internal loop. Two additional simulations, including the MD_A14U simulation with A14U substitution and the MD_A14G_U4C simulation with A14G and U4C substitutions (Table 1), did not show disturbance of the cross-strand stacking, although instabilities of substituted base pairs were detected (data not shown). In the latter simulation, we initially attempted to introduce a canonical G14C4 base pair, but this base pairing was not stable in this context.

Elevated Temperature and No-Salt Simulations of H40. Both simulations performed at 400 K led to disruption of the whole structure within the first 10 ns (data not shown). During the first 5 ns of the no-salt simulation, the major groove rapidly widened to 18 Å, and changes occurred in the internal loop. In particular, the sheared A/G pair and the rH pairs disrupted after 120 and 95 ns, respectively; however, the A5A14 cross-strand stack remained stable, indicating that this stack has considerable stability.

Geometry of the Solution Structure Is Stable in MD Simulations. The MD_NMR_99 simulation of the NMR solution structure is significantly more stable than that of the X-ray structure (see low RMSD value in Table 1). The major groove remained as wide as in the experimental structure, and the interphosphate distances oscillated around the starting values (Figure 4C). The sheared A/G and A/A pairs were stable. It must be admitted that the experimental structure was refined with the same force field, albeit using 500 K in vacuo annealing instead of using explicit solvent at 300 K.

The peculiar *c*WS A/U base pair that is incomplete in the NMR structure exhibits interesting behavior. Its presence conflicts with secondary structure predictions that would place a Watson−Crick base pair there. In addition, the Watson−Crick A−U base pair would be achievable from the incomplete *c*WS starting geometry by a modest rearrangement. This base pair may be essential to understanding the internal loop. The simulation reveals that the incomplete *c*WS base pair is in fact water-bridged, as its sugar−base A15(N1)−U4(O2′) interaction is mediated by water, a substate not apparent from the classification by Leontis and co-workers.[6] The bridging water molecules do not show anomalously long residency times[82] and exchange typically on the time scale of hundreds of picoseconds. Further development of the trajectory revealed two alternative geometries. In the 0−47 ns and 130−160 ns time periods, it was seen in the starting geometry, but it assumed a standard Watson−Crick (*c*WW) conformation in the rest of the simulation. It never sampled geometry with a fully completed *c*WS base pair with direct A15(N1)−U4(O2′) H bond. Therefore, the simulation appears to be consistent with the unusual experimental structure, albeit perhaps subtly biased toward canonical pairing. Note that very small bias of the force field (in terms of free energy) would be sufficient to change the balance between these two substates if they are close in energy. Thus, the simulation behavior does not indicate any large imbalance of the force field.

In the MD_NMR_restr simulation, we imposed a direct A15(N1)−U4(O2′) H bond in the *c*WS A/U base pair via a restraint. At 10 ns, the restraint was released, after which the H bond changed immediately to the water-mediated bond. This is another indication of the overall (qualitative) correctness of the force field, which clearly eliminates the fully paired *c*WS structure, in line with the experiments. This reflects very good balance of the AMBER force field for stacking interactions and noncanonical RNA base pairing.[83,84] At 17 ns, the base pair adopted standard *c*WW geometry, which was stable until the end of the simulation.

MD Simulations of H40 with parmbsc0 Are Similar to the Simulations with parm99. The 100 ns control simulations of the ribosomal H40 run with the parmbsc0 force field[44] (Table 1) showed similar albeit reduced widening of the groove relative to that of the corresponding parm99 simulations (cf. Figure 4B). However, the A5A6 and A5A14 stacks were not disrupted in three out of four simulations (cf. Figure 4B, top). Full details are given in the Supporting Information. The control simulation of the solution structure with parmbsc0 (MD_NMR_bsc0, see Table 1) provided an identical picture to that of the MD_NMR_99 simulation (Supporting Information and Figure 4C).

MD Simulations of H40 in Excess of KCl. In the simulations carried out with an excess of KCl (MD_Ec_K1−3, Table 1), instabilities in base pairing were mostly seen for the U(N3)−A(N7) H bond of the rH pair, in agreement with the parm99 Na^+ simulations (Table 2). However, the two stacks (A5A6 and A5A14) were stable similarly to the parmbsc0 force field simulations.

In the MD_Ec_K1 simulation (KCl, Dang's parameters and parm99, see Table 1), the opening of the major groove was reduced by ∼4 Å compared to the parm99 Na^+ simulation (Figure S5, Supporting Information). In the MD_K2 simulation (KCl, Dang's parameters, parmbsc0) and in the MD_K3 simulation (KCl, Joung's parameters and parm99), the widening of the major groove coupled with disruption of the BPh G13(N1)−A5(O2P) contact was only seen during the first 18 and 30 ns, respectively (Figure S5, Supporting Information). After that, we observed narrowing of the major groove and restoration of the X-ray BPh H bond, in a form of the bifurcated G13(N1, N2)−A5(O2P) 4BPh interaction. This H bond was seen in the MD_Tt_99

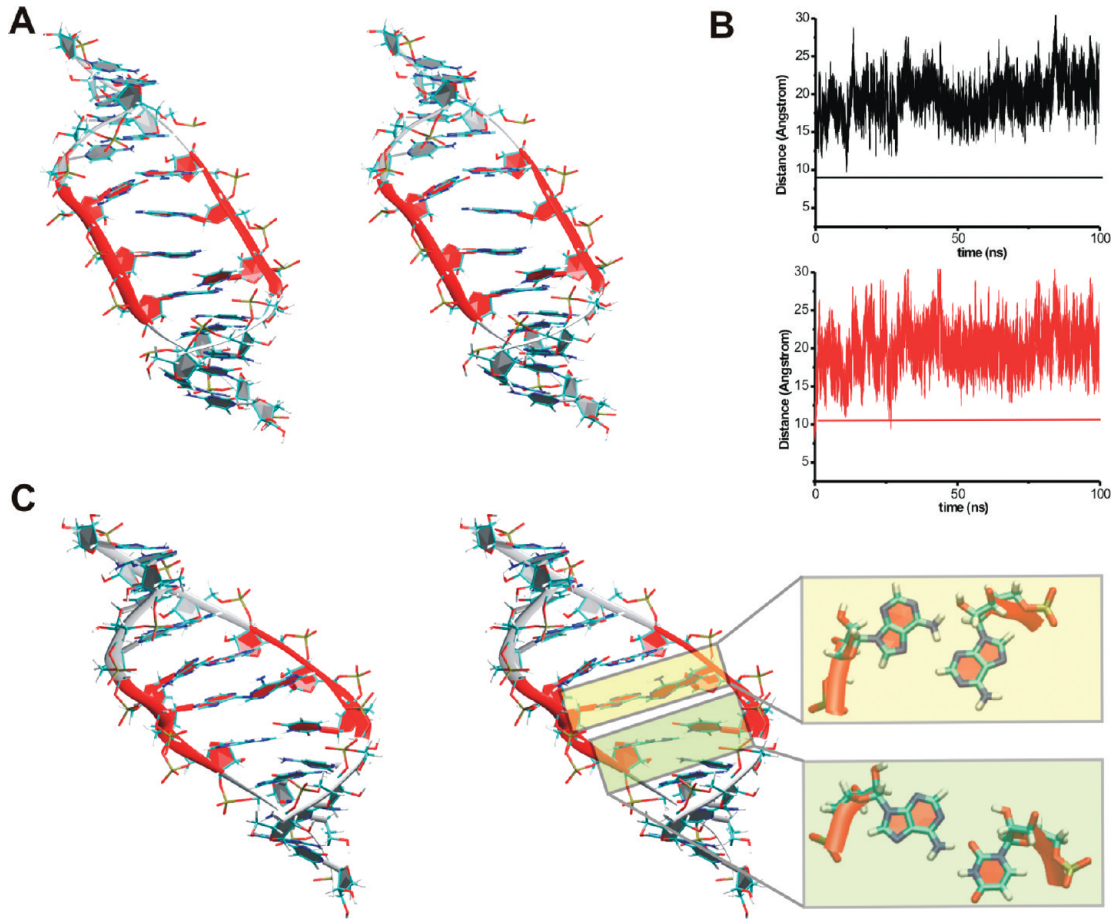


Figure 6. (A) Stereo view of the snapshot structure of the H68 UAA/GAA internal loop from the MD_68 simulation at 5 ns. The structure exhibits a wider (more open) major groove compared to the original geometry (see Figure 1B). (B) Time courses of two interphosphate distances (12P−3P in black and 11P−4P in red) along the MD_68 simulation. (C) Stereo view of the snapshot structure of the H68 UAA/GAA internal loop from the MD_68 simulation at 40 ns with formed sheared A14/A5 and *c*WS A15/U4 pairs, which are highlighted in the color transparent boxes. This structure closely resembles the NMR structure.

simulation (see above). In addition, in the MD_Ec_K3 simulation, the structure expelled the unpaired A5 base from the stem at 23 ns, which was accompanied by subsequent stacking of the sheared pair and the rH pair. A similar event has been detected in the net-neutralizing Na^+ simulation with A5U mutation (see above).

The MD_NMR_K simulation provided a picture close to identical to the MD_NMR_99 simulation. Similarly to the MD_NMR_bsc0, subtle compaction of the major groove by ∼1 Å compared to the parm99 result was detected (Figure S5, Supporting Information). The sheared A/G and A/A base pairs were stable. The A/U base pair showed the starting geometry until 70 ns, while after 70 ns it converted to the canonical *c*WW conformation.

MD Simulation of the H68 X-Ray Structure Converts to the Solution Structure. In the simulation of the ribosomal H68 (MD_H68), the widening of the major groove was also detected (Figures 1B right and 6A), similarly to the H40 simulations (see above and Figure 4).

The interphosphate distances quickly increased up to 20 Å and then fluctuated around this value (Figure 6B). At the beginning of the simulation, a single H bond (A(N6)−G(N7)) formed between A6 and G13 bases, and it was stable until the end of the simulation. This pairing does not correspond to any established base pair family.[6] Around 30 ns, the middle stacking adenines A14 and A5 formed a sheared pair, and the A15 and U4 bases formed a *c*WS pair with one direct bond (A(N6)−U(O2)) and one water-mediated bond (A(N1)−U(O2′); Figure 6C). Both of these pairs occur in the solution structure (Figure 1C). The sheared pair was stable by the end of the simulation, while the *c*WS pair alternated with the *c*WW geometry, similarly to the simulations of the solution structure (see above). The *c*WS geometry was seen in the time periods of 30−45 and 52−57 ns, and the *c*WW geometry in the time periods of 45−52 and 57−100 ns. Importantly, the final transformed H68 geometry is very close to the solution structure, with a RMSD of only 1.6 Å (Figure S6, Supporting Information). The solution structure A6/G13 base pair was, however, not formed. This is due to the fact that G13 in the X-ray structure is in unusual syn conformation. The simulation was not long enough to flip the G13 to the anti conformation, which would lead to entire agreement with the NMR structure. In fact, it cannot be ruled out that the initial G13 syn conformation is an experimental refinement error. In the RNase P RNA X-ray structure, the equivalent guanine is indeed in anti orientation. We have attempted three 150 ns simulations (two with parm99 and one with parmbsc0, data not shown) where the G13 was

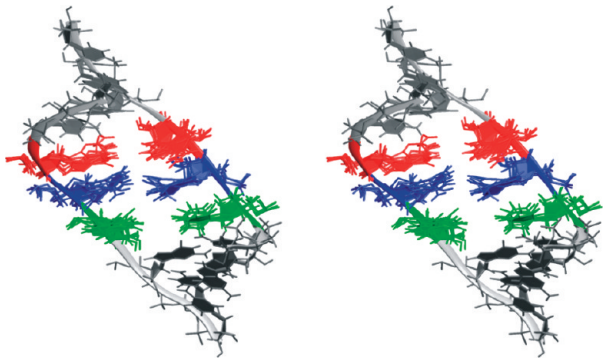


Figure 7. The 3D stereo view of a snapshot of the UAA/GAA structure from the LES_Ec simulation with multiple copies of nucleotides in the internal loop (LES region). The structure resembles the solution structure (Figure 1C); that is, it has a wide major groove and coplanar A6 and G13 (red), A5 and A14 (blue), and U4 and A15 (green) bases.

initially flipped to anti. These simulations, however, did not reveal larger transitions toward the solution structure. It may reflect both a stabilizing effect of the initial G13 anti conformation as well as sampling limitations. Despite this, we still consider the above-analyzed MD_H68 simulation as a solid piece of evidence of the tendency of the X-ray structure to convert spontaneously to the solution structure, which is also supported by the free energy computations (see below).

LES Simulations of the Ribosomal H40. The aim of the LES simulations was to achieve transition from the X-ray H40 structure to the NMR structure, which did not occur in multiple standard simulations. The solution structure internal loop arrangement was sampled only in two LES simulations, which are described below. Other LES simulations are presented in the Supporting Information.

In the first 12 ns of the LES_Ec simulation (Table 1), the original X-ray base pairing of the internal loop was disrupted and the major groove width increased to ∼20 Å. In the 12−40 ns time period, the "multiplied" bases of the internal loop (nucleotides involved in LES) adopted various rapidly changing arrangements and did not form stable base pairs. At 41 ns, G13 and A6, A14 and A5, and A15 and U4 became coplanar (Figure 7), which markedly resembled the solution structure (Figure 1C).

However, the LES bases failed to establish stable pairs. This "NMR-like" arrangement was maintained in the rest of the simulation except for several short disruptions. We started standard MD simulation from this "NMR-like" geometry (see Table 1). After 40 ns, single H bonds formed between G13 and A6, and between A14 and A5; however, the A15 and U4 bases were expelled from the stem. After 70 ns, the internal loop was disrupted, resulting in disturbing of the whole structure (data not shown).

During the first 5 ns of the LES_Dr simulation (Table 1), internal loop base pairs were disrupted and the width of the major groove increased to ∼20 Å, similar to the LES_Ec simulation. At 7 ns, the bulging A15 base flipped into the stem, and during the 10−12 ns time period, the LES bases formed an arrangement where G13 and A6, A14 and A5, and A15 and U4 were coplanar, like in the LES_Ec simulation (Figure 7). In the 12−40 ns time period, the LES bases sampled various unstable arrangements (data not shown). Thus, in summary, LES may show some signs of the transition, but no complete transition was achieved.

Free Energy Calculations. Figure 8 summarizes the MM-PBSA free energy calculations for the parm99 simulations of the UAA/GAN internal loop X-ray and NMR structures.

For H40 simulations, the initial rapid expansion of the major groove (Figure 4) is accompanied by a free energy drop of about 3−6 kcal/mol (Figure 8). The total free energy time course of the H68 loop simulation revealed two marked decreases of free energy. The first one (by ∼10 kcal/mol) can be seen after the first 5 ns, and it corresponds to the rapid expansion of the major groove (Figure 6B), similarly to the H40 total free energy time courses. The second one (by ∼12 kcal/mol) can be seen around 30 ns and it corresponds to the transition of the 2D structure (i.e., formation of the sheared A/A and *c*WS A/U base pairs, see above).

Comparing the averaged total free energies for the 1−100 ns time periods, the NMR structure is predicted to be more stable than the X-ray H40 structure by about 17 kcal/mol (*E.c.* H40 simulation), 19 kcal/mol (*H.m.* H40), 11 kcal/mol (*D.r.* H40), and 18 kcal/mol (*T.t.* H40). The solute entropic term favors the X-ray structure by ∼4−5 kcal/mol, which is consistent with the expectation that the X-ray structure is intrinsically less rigid in isolation. The H68 after the transition (since ∼30 ns) is on average by ∼5 kcal/mol more stable than the NMR structure. With exclusion of the entropic term, the averaged free energies of final H68 and NMR structures would be identical. In summary, the free energy computations give a clear hint that the NMR structure of the UAA/GAN internal loop is indeed intrinsically more stable than the X-ray H40 and H68 structures, albeit the energy difference is probably overestimated as usual with this kind of highly approximate free energy calculation. Note that the free energy computations should in no case be taken quantitatively, despite abundant such attempts in the contemporary literature.

We have further tested an alternative potential of mean force method of free energy computations which was recently used for 16S ribosomal decoding bases 1492 and 1493.[85] We could not use it for the transition between X-ray and solution structures, as we did not see a full transition. However, we used the method to investigate the free energy basin around the H40 X-ray structure while comparing sampling with the parmbsc0 and parm99 force fields. The results are in full detail in the Supporting Information.

NEB and TMD Reveal a Possible Pathway for the Transformation between the NMR and X-Ray H40 Structures. The NEB calculations provide a 32-image pathway where the end points are fixed conformations, with the first image being the energy-minimized NMR structure (Figure 9A) and the final structure being the X-ray structure.

Observations of all pathways reveal that they pass through similar intermediates. Potential energy profiles for the pathways also are similar (Figure S7, Supporting Information). The pathways involve a particular order of structural events. First, A14 breaks its pairing with A5 and moves away from being stacked with A6 and A15 (Figure 9B). A15 slides

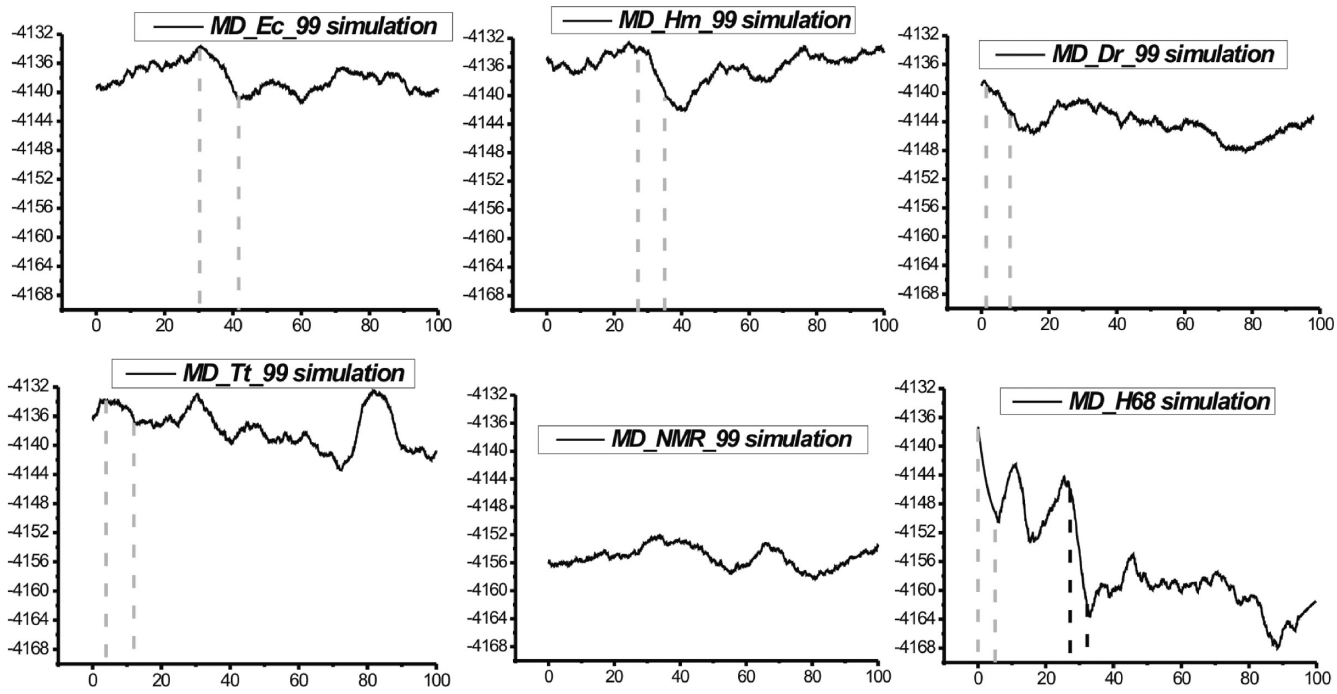


Figure 8. Total free energy time courses in standard net-neutralizing Na^+ simulations with the parm99 force field. The *x* axes stand for time (in nanoseconds), while the *y* axes stand for total free energy (in kcal/mol). The gray vertical dashed lines mark the time period when initial opening of the major groove was observed. In the MD_Tt_99 simulation, the major groove oscillated back and forth with the gray lines indicating the first opening. In the time course of the MD_H68 simulation, the vertical black dashed lines indicate the time period when the sheared A/A and *c*WS A/U pairs formed.

rapidly out of the helix to become the bulged A15 observed in the X-ray structure. A14 moves to pair with U4 as soon as A15 is out of the way and stacks with G16 (Figure 9C). A5 is left unpaired and loses its stacking interaction with U4 to end up hovering over the pairing region of the U4−A14 pair. A5 also shifts further from G13 to become stacked with A6 (Figure 9D) and forms a hydrogen bond with the backbone of the opposite strand, as it is no longer base-paired.

The potential energy profile is a plot of the potential energy for each of the 32 images along the pathway (Figure S7, Supporting Information). The potential energy difference between the product and reactant structure, 87 kcal/mol, is large compared to the above-noted free energy differences. Note that the NEB potential energy does not include the entropic effect of conformational freedom, and it would require a sampling method such as umbrella sampling to relate the NEB potential energy to free energy.

The NEB calculations were run from the NMR structure, which turned out to have the lowest potential energy, to the high-energy X-ray structure. All pathways start with a slight increase in potential energy to about −4060 kcal/mol and remain there until the last few images, where there is a sudden increase in energy to the X-ray structure. There is variation in the slight peaks and valleys throughout the region where the potential energy is about −4060 kcal/mol, but no clear or consistent transition states or intermediates are observed for the 21 NEB pathways. In summary, the limited variation between NEB trials suggests that the conformational change occurs using a single predominant pathway (Figure 9). There is some minor energetic variation as the balance of molecular forces is slightly different for the different pathways, as indicated by the occasional and inconsistent energy minima and maxima in the potential energy profile.

The conformational transitions between NMR and X-ray structures obtained from the TMD simulations are basically identical to the NEB data (Figure 9). In the 20-ns-long MD_TMD_1 simulation, the conversion started at 11 ns (Figure S3, Supporting Information) with disruption of the sheared A14/A5 pair, similar to the NEB result, followed by disruption of the *c*WS A15/U4 pair. Then, the A15 base bulged out of the helix, and the A14 base moved by one base in the strand and created a pair with U4. The A5 base remained unpaired and formed a stack first with G13, and then it moved and stacked with A6. In the MD_TMD_2 simulation, the transition started directly with disruption of the A15/U4 base pair. Otherwise, the transition was identical to the MD_TMD_1 simulation. Additional control simulations revealed identical pictures to the MD_TMD_1 and MD_TMD_2 simulations (data not shown). The energy profiles extracted from the MD_TMD_1 and MD_TMD_2 simulations indicate that the X-ray geometry of the H40 UAA/GAN internal loop has about 30 kcal/mol higher free energy than the NMR structure (Figure S3, Supporting Information). However, the extracted energies must be taken with care because step changes in the RMSD profiles (mainly in the MD_TMD_2 simulation) can be seen (Figure S3, Supporting Information). This indicates insufficient overlap between windows in these regions, which may bias

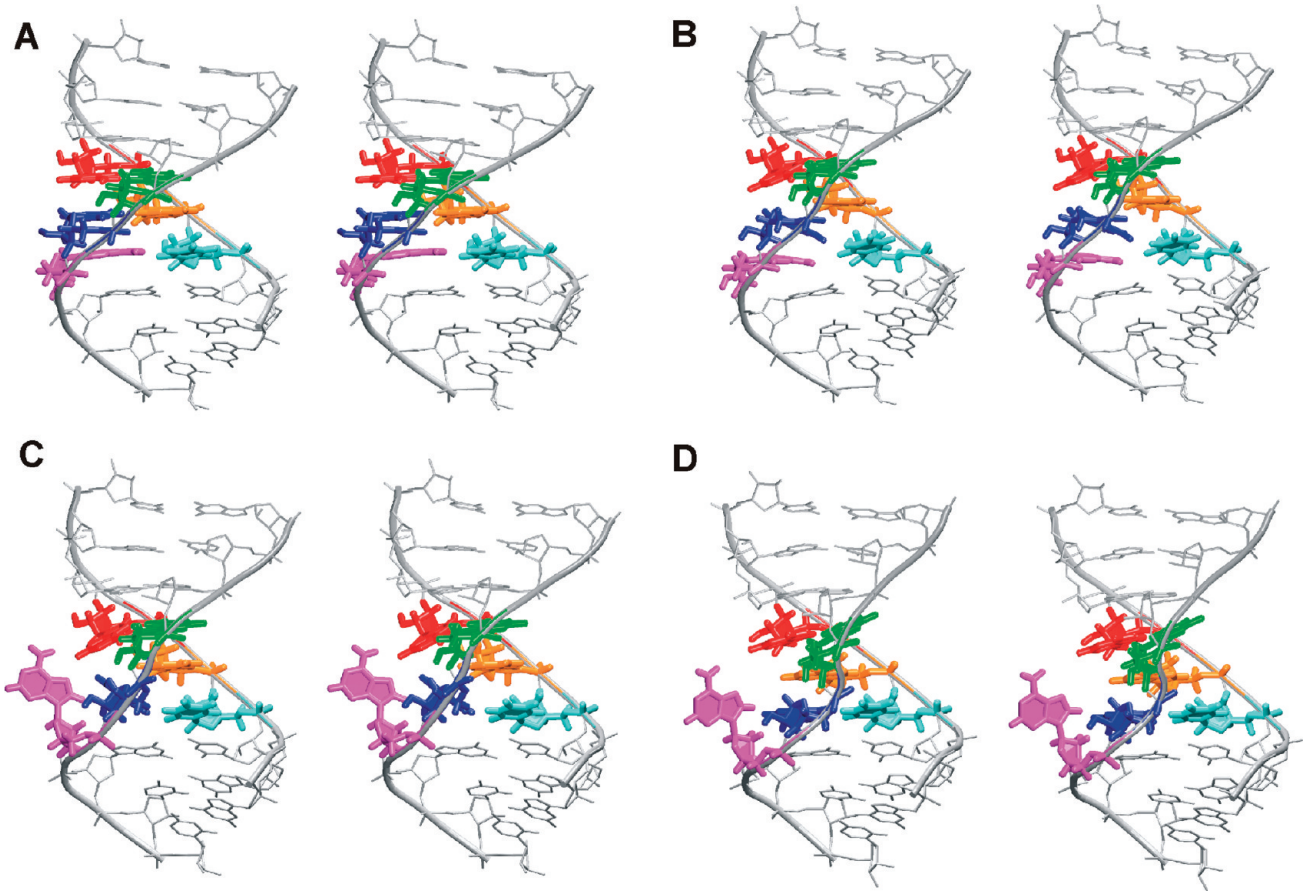


Figure 9. The 3D stereo view of structures showing the conformational transition between NMR and X-ray structures predicted by NEB calculations. G13 is highlighted in red, A6 in green, A14 in blue, A5 in yellow, A15 in magenta, and U4 in cyan. (A) Starting NMR structure. (B) Intermediate structure where the A14/A5 pair breaks and moves away from being stacked with A6 and A15. (C) Intermediate structure where A15 slides out of the helix to become bulged out. (D) Final structure where A14 pairs with U4 while A5 is unpaired and stacks with A6. This structure corresponds to the arrangement of the H40 X-ray structure.

the potential.[86] The trend in the free energy is, however, entirely consistent with the MM-PBSA and NEB data.

Discussion and Conclusions

We employed MD methods to investigate the highly conserved UAA/GAN internal loop of 23S rRNA H40 (Figure 1),[5] which also occurs in six other 23S rRNA helices and in other RNAs.[5] It consists of rH U4/A14 and sheared A6/G13 base pairs interconnected via an unpaired A5 base by two stacks, the A5A14 cross-strand stack and the A5A6 intrastrand stack, as well as a bulged N15 base (Figure 1). The UAA/GAN internal loop in the ribosomal structure has a narrow major groove and wide minor groove (Figure 1). This functional conformation is involved in tertiary contacts with surrounding ribosomal elements that drastically rearrange the base pairing and stacking of the loop compared to its solution structure.[13] These contacts include involvement of conserved adenines in A-minor interactions (Figure 3).

The solution structure contains three noncanonical base pairs (the A6/G13 sheared pair, the A14/A5 sheared pair, and the *c*WS A15/U4 pair) with no unpaired base. The *c*WS A15/U4 base pair observed in the NMR structure is surprising because secondary structure predictions posit a canonical A−U base pair at this position. In addition, the *c*WS A15/U4 pair is incompletely paired.

We studied the H40 UAA/GAN internal loop taken from available X-ray bacterial and archaeal 50S subunits along with the NMR solution structure[13] (Table 1, Figure 1, and Figures S1 and S2, Supporting Information). Furthermore, we investigated the less frequent UAA/GAN geometry from 23S rRNA H68, which adopts yet another (alternative) 2D/3D arrangement (Figure 1). Therefore, the UAA/GAN internal loop is an RNA molecular switch that has functional geometry in folded RNAs that differs from its optimal geometry in isolation.

H40 UAA/GAN Basic Simulations. Unrestrained explicit solvent MD simulations revealed relaxation of the X-ray H40 UAA/GAN internal loop on a scale of tens of nanoseconds. In particular, considerable expansion of the major groove width from the original value of 6−8 Å up to 16−22 Å was detected, coupled with disruption of the X-ray base−phosphate interaction across the major groove (Figures 2 and 4). Further, widening of the major groove was accompanied by replacement of the X-ray A5A6 intrastrand stack by a new A5G13 stack (Figure 5). The newly formed stack was probably more compatible with the wide major groove than the original X-ray one. The relaxed X-ray geometry (Figure 4A) partially resembles the solution structure with its wide major groove with a width of ∼17 Å (Figures 1 and 4). However, the open conformation of the solution structure

may be a result of the NMR refinement procedure, which utilizes NMR-restrained molecular dynamics and energy minimization[13] because the NMR experiment does not provide precise structural information about the sugar phosphate backbone. Further, all standard simulations of the X-ray H40 UAA/GAN exhibited opening events and fluctuations of base pairs in the internal loop (Table 2). Irreversible disruptions of H bonds were also detected in these pairs (Table 2). The increased dynamics of the sheared A/G pairs can be because of the large number of potential hydrogen bonds which cannot all be made simultaneously, which was proposed in an NMR study of GNRA hairpin loops.[87] Despite such dynamics and changes, the overall X-ray H40 secondary structure of the internal loop was maintained in the simulations; that is, the base pairing does not spontaneously rearrange toward the solution structure. Occasionally, in some long simulations, the structure was ultimately disrupted.

The cross-strand A5A14 stack shows only modest fluctuations in our simulations in comparison with the intrastrand A5A6 stack and may be one of the key stabilizing elements of the X-ray secondary structure. It has been suggested[13] that the cross-strand stack allows base pairing of A14 and U4 and additionally compensates for H bonds lost between A15 and U4 and between A14 and A5. We attempted to disrupt the cross-strand stack in the simulations with several mutations (A5U, A14U, and A14G together with U4C, Table 1). The first substitution led to expulsion of the U5 from the stack and subsequent substantial rearrangements, hinting at the key role of A5 not only for the tertiary interactions but also for the stability of the functional X-ray structure of the UAA/GAN loop. The other substitutions had an inconclusive impact on the simulations. Likewise, simulations at elevated temperature and assuming no-salt conditions did not provide any insights into the properties of the UAA/GAN internal loop.

H68 Simulations. In the standard simulation of the ribosomal H68 UAA/GAA loop, we observed a large spontaneous transition clearly toward the solution structure (Figure S6, Supporting Information), except that the A6 and G13 bases did not form any classified base pair, which probably relates with the initial syn orientation of the G13 base. The simulation was not able to overcome this initial syn orientation.

Solution Structure. Simulations of the solution UAA/GAN loop structure were stable (Table 1). The wide major groove remained unchanged, and the base pairs of the loop were stable except for fluctuations of the A15/U4 base pair. In particular, the experimental *c*WS A15/U4 base pair is stabilized by only one direct H bond, despite the fact that two direct bonds are assumed by standard classification.[6] The simulations reveal that there is an additional stabilizing interaction in this base pair, namely, a sugar−base water bridge. In the simulations, this partially paired base pair alternates with the canonical (*c*WW) geometry expected from thermodynamic considerations,[88,89] but not indicated by the NMR experiment.[13] On the other hand, a fully paired *c*WS A15/U4 base pair never formed in the simulations and was immediately disrupted even when initially imposed by restraints, entirely in agreement with NMR. Thus, simulations of the NMR structure of the UAA/GAN internal loop indicate a satisfactory performance of the simulation force field, albeit the balance of the simulation might be subtly shifted toward formation of the canonical A15−U4 base pair. We suggest that the solution structure of the UAA/GAN internal loop represents an interesting test molecule for verification of simulation methods and force fields.

Structural Plasticity of the UAA/GAA Internal Loop. To obtain additional insights, we applied a range of auxiliary methods that can enhance the capabilities of standard simulations. Note, however, that all of these methods necessarily introduce additional approximations and are thus inherently less reliable than standard simulations. The LES technique was applied to enhance sampling of bases in the X-ray H40 loop. All of the LES simulations revealed widening of the major groove and disruption of the internal loop. The internal loop adopted various arrangements where bases involved in the LES region mutually stacked or formed temporary contacts. However, LES was not robust enough to converge into a stable prevalent conformation. Occasionally, the bases formed a secondary structure arrangement similar to the solution conformation (Figures 1 and 7), although no stable base pairing was established.

Total free energies were extracted from the standard simulations utilizing the MM-PBSA method. Free energy time courses along the UAA/GAN H40 X-ray loop trajectories showed a ∼3−6 kcal/mol free energy improvement during the significant 10−15 Å increases in the major groove width (see Figures 4 and 8). Furthermore, MM-PBSA data predicted the free energy of the X-ray H40 UAA/GAN internal loop structure to be ∼10−20 kcal/mol less favorable compared to the NMR structure. A previous experimental study[13] predicted the internal loop of the NMR structure to be favorable by ∼5 kcal/mol when compared to the X-ray H40 functional structure. This estimate was based on an experimental measurement of free energy of the internal loop in the solution structure and a prediction of free energy of the X-ray ribosomal internal loop utilizing a nearest neighbor model.[89,90] Thus, the free energy calculations identify the correct trend but do not reach quantitative accuracy.

The conformational transition between NMR and H40 X-ray structures was investigated by the NEB and TMD methods. Results are mutually consistent. The conversion could start with breaking the sheared A14/A5 pair. This is in accord with the NMR study,[13] which suggested structural dynamics of the A5 base and proposed that the dynamics may provide a pathway for conformational conversion. The transition continues with disruption of the *c*WS A15/U4 pair and bulging out of the A15 base, and eventually formation of the rH U5/A14 pair, the A5A6 stack, and the cross-strand A5A14 stack (Figure 9). Calculations thus predict a likely mechanism for rearrangement of the solution conformation into the functional "ribosomal" X-ray geometry. In the ribosome, the conversion could be induced by an adjacent rRNA (hairpin structure between H39 and H40) and ribosomal protein L20, which binds *E.c.* 23S rRNA at an early stage of ribosomal assembly.[91] The presence of two single H-bond pairs, the sheared A/A and *c*WS A/U pairs, in the

internal loop of the solution structure suggests that the loop structure may be internally weak and easily disrupted by external forces. The weakness of the pairing in the solution structure, which is the global minimum of the UAA/GAN internal loop, may be one of the important prerequisites for its smooth transition to the functionally important substate. This may contribute to determination of the consensus sequence of the UAA/GAN internal loop whose sequence signature is otherwise primarily determined by the X-ray architecture and the tertiary interactions it is involved in.

Force Field Choice. We see a modest difference between the parm99 and parmbc0 force fields, but this does not affect any key conclusions of the paper. Overall, the control MD simulations carried out with the parmbsc0 force field[44] provided a similar picture as the simulations run with parm99[43] (400 ns of comparable trajectories for the H40 initial structure). Nevertheless, in simulations of the ribosomal H40, the major groove width was reduced by 2−4 Å when using parmbsc0 compared to parm99 simulations (Figure 4). In the simulation of the solution structure, the major groove width was reduced only by 1 Å. There are two competing α/γ backbone substates, canonical geometry and t/t conformation (two established A-RNA families, 20 and 24).[92] During the 100 ns portion of the *E.c.* parm99 simulation, we detected a 17% population of nucleotides in the α/γ t/t conformation while the t/t flips are reversible. This is comparable with the 10−15% population reported in the MD study of 16S rRNA H44 and canonical A-RNA.[36] In contrast, the γ-trans states are fully suppressed using the parmbsc0 force field. The suppression of the γ-trans substates allows the major groove narrowing. As the α/γ t/t substate occurs occasionally also in experimental structures and is compatible with overall A-RNA topology, both force fields have satisfactory performance, despite the above-noted difference. The actual propensity of A-RNA to populate the α/γ t/t substates is likely in between the parm99 and parmbsc0 propensities, while the overall difference between the two force fields is, for the present RNA system, small.

In the excess salt KCl simulations, the major groove width was also reduced compared to the net-neutralizing Na^+ simulations (by ∼2−4 Å for the H40 structure). In addition, in some KCl simulations, the major groove after the initial widening returned to the closed X-ray conformation stabilized by the BPh interaction. We suggest that the results are explained by better screening of the phosphates with a higher ionic strength, which allows their closer approach across the groove. Thus, the stability of the functional H40 conformation may be affected by ionic conditions or other interactions reducing the interphosphate repulsion.

Which of the force field options is better? The answer is not unambiguous. The simulations clearly show a tendency of the major groove to widen for the UAA/GAA internal loop when it is taken out of its ribosomal context. We earlier reported widening of the major groove width for the 5S rRNA loop E in parm99 Na^+ simulations. Loop E is an internal loop with seven consecutive noncanonical RNA base pairs. Shorter, ∼10 ns, trajectories gave increases in the P−P distances for the 5S rRNA loop E system by only a few angstroms compared to the X-ray structure.[82] In contrast to H40, however, loop E is in its global minimum in its X-ray structure. All of these results may give an impression that simulations tend to overshoot the RNA major groove width. This effect is larger with parm99 than with parmbsc0 and can be reduced by using excess salt conditions instead of net-neutralization. Our very recent reference simulations on canonical A-RNA,[93] however, show that the picture is more complex. These reference simulations also usually show a tendency for widening of the major groove in simulations compared to the X-ray structures, which is larger with the parm99 than with the parmbsc0 force field. These simulations, however, also show that the A-RNA major groove width and its relaxation depend on the base sequence and are different for different X-ray structures. Therefore, there is no unambiguous experimental target value of the major groove width, as the experimental values depend on the sequence and crystallization conditions. In some cases, parm99 with net-neutralization can remain closest to the experimental structures. Therefore, the right interpretation is that the RNA major groove width is sensitive to the sequence, environment and molecular interactions and the simulations reflect this groove width plasticity. We suggest that all force field options utilized in the present study are justified for RNA simulations. Nevertheless, RNA simulations can perhaps indirectly profit from the parmbsc0 force field choice, since its tendency to keep the major groove more closed may reduce the likelihood of irreversible structural disruptions during some large temporary major groove width fluctuation events. On the other hand, the parmbsc0 force field appears to somewhat rigidify the simulated structures compared to parm99, as evidenced by less frequent changes of the adenine stacks (see above). We cannot tell, however, whether this behavior is an improvement or not compared with the parm99 force field.

The H40 UAA/GAA simulations further show some local instabilities, and some long simulations result in an entire loss of the X-ray structure topology for the UAA/GAA internal loop without approaching closer to the solution structure. Development of the simulations on a much longer time scale is thus uncertain. The perturbation may either reflect the genuine internal instability of the functional (ribosomal) structure of UAA/GAA when considered in isolation or it may be an accumulation of force field imbalances in our long simulations. Most likely, both factors contribute.

Concluding Remarks. The simulations provide atomistic characterization of the structural dynamics of the UAA/GAA internal loop in three distinct experimental topologies. The simulations appear to be consistent with experimental data and give new insights. Our study is probably the first simulation study of a recurrent RNA non-Watson−Crick element that is not autonomous; that is, it folds only in specific contexts. The H40 simulations do not spontaneously transform to the solution (ground state) structure, and such transition is probably beyond the limits of contemporary computational chemistry. However, almost complete transformation was seen for the alternative H68 X-ray structure. Methods like TMD or NEB can achieve transformation between the H40 and solution UAA/GAA topologies; however, they impose an artificially selected path and require

a priori knowledge of both the starting and final structures. Free energy computations can provide some very crude estimates of the free energy trends but are probably far from reaching even qualitative accuracy.

The results suggest that the H40 and H68 internal loops are under stress due to tertiary and quaternary interactions, and that H68 can relax to its conformation in isolation much faster than H40 if the interactions with its surroundings are relieved or altered. Thus, the MD results suggest that the different structures induced by tertiary and quaternary interactions may also have implications for the temporal control of events. Both the MD and NMR results indicate that there is no significant population of higher free energy structures in isolated RNA. This suggests that the approach of other parts of the rRNA or of protein induces a conformational change rather than trapping a minor species. This type of conformational switch may be important for assembly or movement in molecular machines such as the ribosome. We demonstrate that, despite the above-explained limitations, modern MD-based computations can complement experimental techniques and provide insights into the role of molecular interactions in shaping RNA building blocks.

Acknowledgment. This work was supported by the Ministry of Education of the Czech Republic [grants MSM0021622413 and LC06030], by the Grant Agency of the Academy of Sciences of the Czech Republic [grant numbers 1QS500040581, IAA400040802, and KJB400040901], and by grants GA203/09/1476 and 203/09/H046, Grant Agency of the Czech Republic. This work was also supported by the Academy of Sciences of the Czech Republic, grant numbers AV0Z50040507 and AV0Z50040702. The work was also partially supported by the United States National Institutes of Health Grant R01HG004002 to D.H.M. and GM22939 to D.H.T. The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement number 205872. We thank Dr. Eva Fadrna for help with the preparation of LES structures.

Supporting Information Available: Detailed description of the standard MD simulations run with parm99, description of the standard MD simulations run with parmbsc0, and description of the LES simulations of ribosomal H40. Simulated annealing protocol used for minimization in the NEB trials (in Table S1). Calculation of the free energy basin around the H40 X-ray structure based on parm99 and parmbsc0 simulations using the potential of mean force coordinate method, supplementary Figures S1−S9. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **2000**, *289*, 905–920.

(2) Harms, J. M.; Schluenzen, F.; Zarivach, R.; Bashan, A.; Gat, S.; Agmon, I.; Bartels, H.; Franceschi, F.; Yonath, A. High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **2001**, *107*, 679–688.

(3) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M. J.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. Structure of the 30S ribosomal subunit. *Nature* **2000**, *407*, 327–339.

(4) Leontis, N. B.; Westhof, E. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.* **2003**, *13*, 300–308.

(5) Lee, J. C.; Gutell, R. R.; Russell, R. The UAA/GAN internal loop motif: A new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions. *J. Mol. Biol.* **2006**, *360*, 978–988.

(6) Leontis, N. B.; Stombaugh, J.; Westhof, E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **2002**, *30*, 3497–3531.

(7) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* **2008**, *56*, 215–252.

(8) Jaeger, L.; Verzemnieks, E. J.; Geary, C. The UA_handle: a versatile submotif in stable RNA architectures. *Nucleic Acids Res.* **2009**, *37*, 215–230.

(9) Schuwirth, B. S.; Borovinskaya, M. A.; Hau, C. W.; Zhang, W.; Vila-Sanjurjo, A.; Holton, J. M.; Cate, J. H. Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **2005**, *310*, 827–834.

(10) Kazantsev, A. V.; Krivenko, A. A.; Harrington, D. J.; Holbrook, S. R.; Adams, P. D.; Pace, N. R. Crystal structure of a bacterial ribonuclease P RNA. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13392–13397.

(11) Nissen, P.; Ippolito, J. A.; Ban, N.; Moore, P. B.; Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 4899–4903.

(12) Guillier, M.; Allemand, F.; Raibaud, S.; Dardel, F.; Springer, M.; Chiaruttini, C. Translational feedback regulation of the gene for L35 in Escherichia coli requires binding of ribosomal protein L20 to two sites in its leader mRNA: A possible case of ribosomal RNA-messenger RNA molecular mimicry. *RNA* **2002**, *8*, 878–889.

(13) Shankar, N.; Kennedy, S. D.; Chen, G.; Krugh, T. R.; Turner, D. H. The NMR structure of an internal loop from 23S ribosomal RNA differs from its structure in crystals of 50S ribosomal subunits. *Biochemistry* **2006**, *45*, 11776–11789.

(14) Auffinger, P.; Bielecki, L.; Westhof, E. Symmetric K^+ and Mg^{2+} ion-binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J. Mol. Biol.* **2004**, *335*, 555–571.

(15) Auffinger, P.; Hashem, Y. Nucleic acid solvation: from outside to insight. *Curr. Opin. Struct. Biol.* **2007**, *17*, 325–333.

(16) Cojocaru, V.; Klement, R.; Jovin, T. M. Loss of G-A base pairs is insufficient for achieving a large opening of U4 snRNA K-turn motif. *Nucleic Acids Res.* **2005**, *33*, 3435–3446.

(17) Deng, N. J.; Cieplak, P. Molecular dynamics and free energy study of the conformational equilibria in the UUUU RNA hairpin. *J. Chem. Theory Comput.* **2007**, *3*, 1435–1450.

(18) Krasovska, M. V.; Sefcikova, J.; Spackova, N.; Sponer, J.; Walter, N. G. Structural dynamics of precursor and product of the RNA enzyme from the hepatitis delta virus as revealed by molecular dynamics simulations. *J. Mol. Biol.* **2005**, *351*, 731–748.

(19) Li, W.; Sengupta, J.; Rath, B. K.; Frank, J. Functional conformations of the L11-ribosomal RNA complex revealed by correlative analysis of cryo-EM and molecular dynamics simulations. *RNA* **2006**, *12*, 1240–1253.

(20) Reblova, K.; Fadrna, E.; Sarzynska, J.; Kulinski, T.; Kulhanek, P.; Ennifar, E.; Koca, J.; Sponer, J. Conformations of flanking bases in HIV-1 RNA DIS kissing complexes studied by molecular dynamics. *Biophys. J.* **2007**, *93*, 3932–3949.

(21) Romanowska, J.; Setny, P.; Trylska, J. Molecular dynamics study of the ribosomal A-site. *J. Phys. Chem. B* **2008**, *112*, 15227–15243.

(22) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C. S. Simulating movement of tRNA into the ribosome during decoding. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15854–15859.

(23) Vaiana, A. C.; Westhof, E.; Auffinger, P. A molecular dynamics simulation study of an aminoglycoside/A-site RNA complex: conformational and hydration patterns. *Biochimie* **2006**, *88*, 1061–1073.

(24) Kopitz, H.; Zivkovic, A.; Engels, J. W.; Gohlke, H. Determinants of the unexpected stability of RNA fluorobenzene self pairs. *Chembiochem* **2008**, *9*, 2619–2622.

(25) Mazier, S.; Genest, D. Insight into the intrinsic flexibility of the SL1 stem-loop from genomic RNA of HIV-1 as probed by molecular dynamics simulation. *Biopolymers* **2008**, *89*, 187–196.

(26) Mokdad, A.; Krasovska, M. V.; Sponer, J.; Leontis, N. B. Structural and evolutionary classification of G/U wobble base-pairs in the ribosome. *Nucleic Acids Res.* **2006**, *34*, 1326–1341.

(27) Rhodes, M. M.; Reblova, K.; Sponer, J.; Walter, N. G. Trapped water molecules are essential to structural dynamics and function of a ribozyme. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 13380–13385.

(28) Schneider, C.; Brandl, M.; Suhnel, J. Molecular dynamics simulation reveals conformational switching of water-mediated uracil-cytosine base-pairs in an RNA duplex. *J. Mol. Biol.* **2001**, *305*, 659–667.

(29) Villa, A.; Wöhnert, J.; Stock, G. Molecular dynamics simulation study of the binding of purine bases to the aptamer domain of the guanine sensing riboswitch. *Nucleic Acids Res.* **2009**, *37*, 4774–4786.

(30) Zhuang, Z.; Jaeger, L.; Shea, J. E. Probing the structural hierarchy and energy landscape of an RNA T-loop hairpin. *Nucleic Acids Res.* **2007**, *35*, 6995–7002.

(31) Razga, F.; Koca, J.; Sponer, J.; Leontis, N. B. Hinge-like motions in RNA kink-turns: The role of the second A-minor motif and nominally unpaired bases. *Biophys. J.* **2005**, *88*, 3466–3485.

(32) Spackova, N.; Sponer, J. Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucleic Acids Res.* **2006**, *34*, 697–708.

(33) Reblova, K.; Razga, F.; Li, W.; Gao, H.; Frank, J.; Šponer, J., Dynamics of the Base of Ribosomal A-site Finger Revealed by Molecular Dynamics Simulations and Cryo-EM. *Nucleic Acids Res.* **2009**, DOI: 10.1093/nar/gkp1057.

(34) Curuksu, J.; Sponer, J.; Zacharias, M. Elbow Flexibility of the kt38 RNA Kink-Turn Motif Investigated by Free-Energy Molecular Dynamics Simulations. *Biophys. J.* **2009**, *97*, 2004–2013.

(35) Razga, F.; Koca, J.; Mokdad, A.; Sponer, J. Elastic properties of ribosomal RNA building blocks: molecular dynamics of the GTPase-associated center rRNA. *Nucleic Acids Res.* **2007**, *35*, 4007–4017.

(36) Reblova, K.; Lankas, F.; Razga, F.; Krasovska, M. V.; Koca, J.; Sponer, J. Structure, dynamics, and elasticity of free 16S rRNA helix 44 studied by molecular dynamics simulations. *Biopolymers* **2006**, *82*, 504–520.

(37) Cheatham, T. E., III; Young, M. A. Molecular dynamics simulation of nucleic acids: successes, limitation, and promise. *Biopolymers* **2001**, *56*, 232–256.

(38) McDowell, S. E.; Spackova, N.; Sponer, J.; Walter, N. G. Molecular dynamics simulations of RNA: An in silico single molecule approach. *Biopolymers* **2007**, *85*, 169–184.

(39) Orozco, M.; Perez, A.; Noy, A.; Luque, F. J. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* **2003**, *32*, 350–364.

(40) Sponer, J.; Lankas, F. *Challenges and Advances in Computational Chemistry and Physics: Computational Studies of RNA and DNA*; Springer: Dordrecht, Netherlands, 2006.

(41) MacKerell, A. D.; Nilsson, L. Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.* **2008**, *18*, 194–199.

(42) Orozco, M.; Noy, A.; Perez, A. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.* **2008**, *18*, 185–193.

(43) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(44) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the amber force field for nucleic acids. Improving the description of α/γ conformers. *Biophys. J.* **2007**, *92*, 3817–3829.

(45) Klein, D. J.; Moore, P. B.; Steitz, T. A. The Roles of Ribosomal Proteins in the Structure, Assembly and Evolution of the Large Ribosomal Subunit. *J. Mol. Biol.* **2004**, *340*, 141–177.

(46) Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **2006**, *313*, 1935–1942.

(47) Gao, Y. G.; Selmer, M.; Dunham, C. M.; Weixlbaumer, A.; Kelley, A. C.; Ramakrishnan, V. The Structure of the Ribosome with Elongation Factor G Trapped in the Post-translocational State. *Science* **2009**, *326*, 694–699.

(48) Schmeing, T. M.; Voorhees, R. M.; Kelley, A. C.; Gao, Y. G.; Murphy, F. V.; Weir, J. R.; Ramakrishnan, V. The Crystal Structure of the Ribosome Bound to EF-Tu and Aminoacyl-tRNA. *Science* **2009**, *326*, 688–694.

(49) Zirbel, C. L.; Sponer, J. E.; Sponer, J.; Stombaugh, J.; Leontis, N. B. Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.* **2009**, *37*, 4898–4918.

(50) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecule. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(51) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani,

F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.

(52) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(53) Aqvist, J. Ion water interaction potentials derived from free-energy perturbation simulations. *J. Phys. Chem.* **1990**, *94*, 8021–8024.

(54) Joung, I. S.; Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(55) Dang, L. X.; Kollman, P. A. Free energy of association of the K+: 18-crown-6 complex in water: a new molecular dynamics study. *J. Phys. Chem.* **1995**, *99*, 55–58.

(56) Smith, D. E.; Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **1994**, *100*, 3757–3766.

(57) Auffinger, P.; Cheatham, T. E.; Vaiana, A. C. Spontaneous formation of KCl aggregates in biomolecular simulations: A force field issue. *J. Chem. Theory Comput.* **2007**, *3*, 1851–1859.

(58) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

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

(60) Zhou, R.; Berne, B. J.; Germain, R. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *18*, 14931–14936.

(61) Csaszar, K.; Spackova, N.; Stefl, R.; Sponer, J.; Leontis, N. B. Molecular dynamics of the frame-shifting pseudoknot from beet western yellows virus: The role of non-Watson-Crick base-pairing, ordered hydration, cation binding and base mutations on stability and unfolding. *J. Mol. Biol.* **2001**, *313*, 1073–1091.

(62) Darden, T.; Pearlman, D.; Pedersen, L. Ionic charging free energies: spherical versus periodic boundary condition. *J. Phys. Chem.* **1998**, *109*, 10921–10935.

(63) Elber, R.; Karplus, M. Enhanced sampling in molecular dynamics: use of the time dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.* **1990**, *112*, 9161–9175.

(64) Roitberg, A.; Elber, R. Modeling side chains in peptide and proteins: application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **1991**, *95*, 9277–9287.

(65) Simmerling, C.; Elber, R. Hydrophobic "collapse" in a cyclic hexapeptide: Computer simulations of CHDLFC and CAAAAC in water. *J. Am. Chem. Soc.* **1994**, *116*, 2534–2547.

(66) Fadrna, E.; Spackova, N.; Stefl, R.; Koca, J.; Cheatham, T. E.; Sponer, J. Molecular dynamics simulations of guanine quadruplex loops: Advances and force field limitations. *Biophys. J.* **2004**, *87*, 227–242.

(67) Simmerling, C.; Miller, J. L.; Kollman, P. A. Combined locally enhanced sampling and Particle Mesh Ewald as a strategy to locate the experimental structure of a nonhelical nucleic acid. *J. Am. Chem. Soc.* **1998**, *120*, 7149–7155.

(68) Fadrna, E.; Spackova, N.; Sarzynska, J.; Koca, J.; Orozco, M.; Cheatham, T. E., III; Kulinski, T.; Sponer, J. Single stranded loops of quadruplex DNA as key benchmark for testing nucleic acids force fields. *J. Chem. Theory Comput.* **2009**, *5*, 2514–2530.

(69) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(70) Srinivasan, J.; Cheatham, T. E.; 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*, 9401–9409.

(71) Luo, R.; David, L.; Gilson, M. K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, *23*, 1244–1253.

(72) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(73) Brooks, B.; Karplus, M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **1983**, *80*, 6571–6575.

(74) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Ross, W. S.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.

(75) Alfonso, D. R.; Jordan, K. D. A flexible nudged elastic band program for optimization of minimum energy pathways using ab initio electronic structure methods. *J. Comput. Chem.* **2003**, *24*, 990–996.

(76) Jonsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccoti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; pp 384−404.

(77) Henkelman, G.; Jonsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(78) Mathews, D. H.; Case, D. A. Nudged elastic band calculation of minimal energy paths for the conformational change of a GG non-canonical pair. *J. Mol. Biol.* **2006**, *357*, 1683–1693.

(79) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(80) Noy, A.; Pérez, A.; Laughton, C. A.; Orozco, M. Theoretical study of large conformational transitions in DNA: the B<->A conformational change in water and ethanol/water. *Nucleic Acids Res.* **2007**, *35*, 3330–3338.

(81) Aci, S.; Mazier, S.; Genest, D. Conformational pathway for the kissing complex - Extended dimer transition of the SL1 stem-loop from genomic HIV-1 RNA as monitored by targeted molecular dynamics techniques. *J. Mol. Biol.* **2005**, *351*, 520–530.

(82) Reblova, K.; Spackova, N.; Stefl, R.; Csaszar, K.; Koca, J.; Leontis, N. B.; Sponer, J. Non-Watson-Crick basepairing and hydration in RNA motifs: Molecular dynamics of 5S rRNA loop E. *Biophys. J.* **2003**, *84*, 3564–3582.

(83) Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. Performance of empirical potentials (AMBER, CFF95, CVFF, CHARMM, OPLS, POLTEV), semiempirical quantum chemical methods (AM1, MNDO/M, PM3), and ab initio Hartree-Fock method for interaction of DNA bases: Comparison with nonempirical beyond Hartree-Fock results. *J. Comput. Chem.* **1997**, *18*, 1136–1150.

(84) Sponer, J. E.; Spackova, N.; Kulhanek, P.; Leszczynski, J.; Sponer, J. Non-Watson-Crick base pairing in RNA. Quantum chemical analysis of the cis Watson-Crick/sugar edge base pair family. *J. Phys. Chem. A* **2005**, *109*, 2292–2301.

(85) Vaiana, A. C.; Sanbonmatsu, K. Y. Stochastic Gating and Drug-Ribosome Interactions. *J. Mol. Biol.* **2009**, *386*, 648–661.

(86) Wood, R. H.; Muhlbauer, W. C. F.; Thompson, P. T. Systematic errors in free energy perturbation calculations due to a finite sample of configuration space: sample size hysteresis. *J. Phys. Chem.* **1991**, *95*, 6670–6675.

(87) Jucker, F. M.; Heus, H. A.; Yip, P. F.; Moors, E. H. M.; Pardi, A. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **1996**, *264*, 968–980.

(88) Mathews, D. H.; Disney, M. D.; Childs, J. L.; Schroeder, S. J.; Zuker, M.; Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7287–7292.

(89) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **1999**, *288*, 911–940.

(90) Xia, T.; Santa Lucia, J., Jr.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **1998**, *37*, 14719–14735.

(91) Nowotny, V.; Nierhaus, K. H. Protein L20 from the large subunit of Escherichia coli ribosomes is an assembly protein. *J. Mol. Biol.* **1980**, *137*, 391–397.

(92) Schneider, B.; Moravek, Z.; Berman, H. M. RNA conformational classes. *Nucleic Acids Res.* **2004**, *32*, 1666–1677.

(93) Besseova, I.; Otyepka, M.; Reblova, K.; Sponer, J. Dependence of A-RNA simulations on the choice of the force field and salt strength. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10701–10711.

CT900440T

JCTC Journal of Chemical Theory and Computation

Hydration of the Lowest Triplet States of the DNA/RNA Pyrimidines

Andrew M. Rasmussen, Maria C. Lind, Sunghwan Kim, and Henry F. Schaefer III*

Center for Computational Quantum Chemistry, University of Georgia, Athens, Georgia 30602

Received September 10, 2009

Abstract: The effects of hydration on the lowest triplet states of the DNA/RNA pyrimidines have been studied by including one and two water molecules explicitly. Three configurations for the singly hydrated cytosine moiety were located, and six for the doubly hydrated system. For thymine and uracil, four singly and eight doubly hydrated structures were found. The singlet−triplet energy gaps of all three pyrimidines (cytosine, thymine, and uracil) fall in the low-energy range of ultraviolet radiation (UVA). Energetic excited states can be a step leading to lesions in DNA, such as a mismatched base pairs. Although the adiabatic and vertical electronic excitation energies for all three pyrimidines slightly increase upon inclusion of additional water molecules, this effect upon the excitation energies is much smaller than hydration effects upon the electron affinities and ionization energies of the three nucleobases. Because both the ground state and the triplet state are neutral, the hydration energy difference between the two states is not significant (compared to those between the neutral and charged species), making the excitation energy less sensitive to hydration.

Introduction

Growing concern for the effects of radiation damage on living cells has motivated the study of various mechanisms of such damage on the molecular scale. Radiative damage to DNA can occur both through direct ultraviolet (UV) exposure, or indirectly, as a result of interaction with reactive products (often reactive oxygen-containing radicals) created by radiation damage to other nearby molecules. Helpful reviews describing the effects of UV light on DNA have recently appeared.[1,2]

An example of a directly UV-induced photoproduct, the cyclobutane pyrimidine dimer (CPD) has been known and studied for nearly half a century.[3] CPDs are the primary photoproducts of DNA UV damage,[1,4,5] and can result from exposure to UVA and UVB light.[6,7] A majority of UV-induced mutations occur due to the formation of CPDs;[5,8] they are mutagenically effective due to their slow rate of repair and high rate of bypass by nucleic acid polymerases[9,10] (the cell's mechanism for repair to damaged DNA), thus allowing their persistence in the genome. Although many vertebrates such as fish, reptiles, and marsupials have an alternative repair mechanism involving photolyases,[11−16] which reverse the CPD lesion to two pyrimidine monomer units using visible light, it is believed that placental mammals including humans do not have this enzyme.[16−19] Formation of CPDs is sequence-specific,[6,20,21] and various combinations of pyrimidines [usually thymine (T) and cytosine (C)] can arise, including T-T, C-T, and C-C dimers. Especially after exposure to UVA radiation, the most common CPD produced is the T-T dimer.[5,6] Various experimental and theoretical studies have shown that it is formed through an ultrafast triplet-energy exchange,[5,22,23] for which a mechanism has recently been proposed.[24] Hence it is important to understand the configurational changes within the pyrimidines that result from excitation, leading to the formation of CPDs and other photoproducts.

The singlet−triplet energy separations (gaps) of the pyrimidines fall well within the UVA absorption range (cytosine: 3.50 eV,[25] uracil: 3.65−3.68 eV,[26,27] thymine: 3.6 eV[28]). For all of the pyrimidines, the lowest-lying excited states are triplets; the excited singlets lie higher energetically (at ∼4 eV).[29,30] Using electron energy loss spectroscopy

* Corresponding author e-mail: sch@uga.edu.

10.1021/ct900478c
Published on Web 02/03/2010

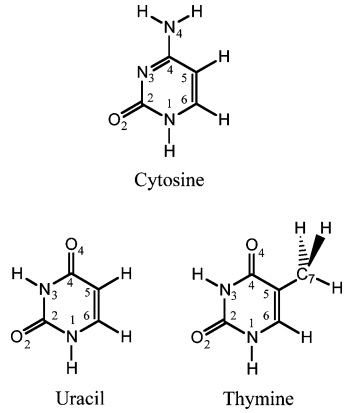


Figure 1. Canonical structures of uracil, thymine, and cytosine with atom numbering schemes.

(EELS), Abouaf, Pommier, and Dunet determined the singlet−triplet energy gap of thymine to be 3.6 (±0.08) eV.[28] The same group later measured the singlet−triplet energy gap of cytosine at 3.50 eV[25] and uracil at 3.65 (±0.05) eV.[31] A recent study by Bosca et al.[32] has also delivered an estimate of the adiabatic triplet excitation energy of thymine in DNA of ∼2.8 eV. Recent DFT computations by Nguyen and co-workers are in good agreement with experimental results.[31] Although hydration plays an important role in biological systems, few studies have explored the hydration effects on the triplet excited states of the DNA/RNA nucleobases.[33]

Much work has been done on the hydration of nucleobases in their ground states. Over the years, many studies have detailed the effects of discrete and continuous hydration of the DNA/RNA nucleobases[34−44] as well as studies focusing specifically on cytosine,[45−58] uracil,[59−73] and thymine.[74−76] In the present work, hydration effects on the lowest triplet states of the three pyrimidine nucleobases have been studied using density functional theory. In particular, structural changes, changes in the triplet excitation energies upon hydration, sites favoring hydration, and hydration energies are reported.

Computational Methods

All computations were performed using the Gaussian 94 computational chemistry software package.[77] Only complexes of the canonical forms of the pyrimidines (Figure 1) were considered, with the water molecules hydrogen bonded in the plane of the molecule. A search was carried out for new monohydration sites for the triplet state of the three bases including nonplanar complexes, in which a water molecule may interact with the aromatic π-system of the pyrimidine ring, but none were encountered.

For all computations, a specially calibrated double-ζ quality basis set with polarization and diffuse Gaussian functions (DZP++) was used. This basis set is constructed with the Huzinaga−Dunning *sp* contractions, adding one set of five *d*-type polarization functions for each C, N, and O atom, and one set of *p*-type polarization functions for each H atom.[78,79] Lee's prescription,[80]

$$\alpha_{\text{diffuse}} = \frac{1}{2}\left(\frac{\alpha_1}{\alpha_2} + \frac{\alpha_2}{\alpha_3}\right)\alpha_1$$

determined the even-tempered orbital exponents ($\alpha_1 < \alpha_2 < \alpha_3$), with the final DZP++ basis set containing six functions per H atom and 19 functions per C, N, and O atom. When tested, the use of a similar triple-ζ quality (TZ2P++) basis set afforded no significant change in results for several times the computational cost, so it was not further employed. Structural optimizations and harmonic frequency analyses for each base and its corresponding mono- and dihydrates were obtained using the B3LYP density functional, a combination of Becke's three-parameter functional (B3),[81] with the correlation functional of Lee, Yang, and Parr (LYP).[82] The self-consistent isodensity polarized continuum model (SCIPCM)[83] was employed to take into account the effect of macrosolvation upon the energetics of the pyrimidine hydrates. The SCIPCM single point energies at their optimized gas-phase geometries were computed at the B3LYP/DZP++ level of theory, with the dielectric constant of water ($\varepsilon = 78.39$) and the isodensity value of 0.0001.

In the present work, we estimated physical properties of interest as follows: the vertical excitation (VEx) energy is

Table 1. Relative Energies (E_{rel}) in kcal mol^{-1} for Singly and Doubly Hydrated Uracil, Thymine, and Cytosine Ground State and Triplet Excited State (Denoted by *)[a]

	E_{rel}				E_{rel}		
structure	gas		SCIPCM	Gas structure	gas		SCIPCM
			uracil hydrate				
U1A	0.00	(0.00)	0.00	U1A*	0.00	(0.00)	0.00
U1B	1.61	(1.52)	0.23	U1B*	1.41	(1.42)	0.36
U1C	2.30	(2.13)	0.68	U1C*	1.91	(1.79)	0.65
U1D	3.38	(3.05)	2.03	U1D*	2.80	(2.49)	1.97
U2A	0.00	(0.00)	0.00	U2A*	0.00	(0.00)	0.00
U2B	2.59	(2.37)	1.14	U2B*	2.11	(2.03)	1.68
U2C	3.40	(2.94)	3.57	U2C*	3.73	(3.23)	3.87
U2D	4.01	(3.55)	2.01	U2D*	3.25	(2.90)	1.71
U2E	4.71	(4.09)	4.14	U2E*	4.61	(3.96)	4.14
U2F	5.44	(4.69)	5.46	U2F*	5.21	(4.50)	5.49
U2G	7.43	(6.47)	6.03	U2G*	6.89	(6.06)	6.04
U2H	7.36	(6.50)	5.77	U2H*	6.81	(6.11)	5.81
			thymine hydrate				
T1A	0.00	(0.00)	0.00	T1A*	0.00	(0.00)	0.00
T1B	1.73	(1.60)	0.28	T1B*	1.42	(1.39)	0.24
T1C	2.12	(1.93)	0.59	T1C*	1.95	(1.81)	0.68
T1D	4.25	(3.80)	2.24	T1D*	3.62	(3.24)	2.07
T2A	0.00	(0.00)	0.00	T2A*	0.00	(0.00)	0.00
T2B	2.80	(2.60)	1.28	T2B*	2.22	(2.19)	1.10
T2C	3.51	(3.03)	3.55	T2C*	3.61	(3.15)	3.63
T2D	3.82	(3.36)	1.92	T2D*	3.38	(3.06)	1.83
T2E	4.52	(3.95)	4.02	T2E*	4.57	(3.98)	4.10
T2F	6.23	(5.42)	5.64	T2F*	5.91	(5.08)	5.53
T2G	8.11	(7.11)	6.11	T2G*	7.70	(6.69)	6.15
T2H	8.51	(7.57)	6.15	T2H*	7.75	(6.96)	5.82
			cytosine hydrate				
C1A	0.00	(0.00)	0.00	C1A*	1.38	(1.27)	1.61
C1B	0.61	(0.54)	0.09	C1B*	0.00	(0.00)	0.00
C1C	6.08	(5.14)	2.94	C1C*	4.96	(4.03)	2.80
C2A	0.00	(0.00)	0.00	C2A*	1.84	(1.68)	2.08
C2B	1.36	(1.24)	0.80	C2B*	0.38	(0.23)	0.23
C2C	1.47	(1.24)	2.09	C2C*	2.54	(2.18)	3.17
C2D	2.78	(2.37)	2.70	C2D*	0.00	(0.00)	0.00
C2E	6.65	(5.68)	4.66	C2E*	7.52	(6.26)	5.84
C2F	7.30	(6.30)	4.86	C2F*	5.96	(4.97)	4.14

[a] Zero-point vibrational energy (ZPVE)−corrected values are in parentheses.

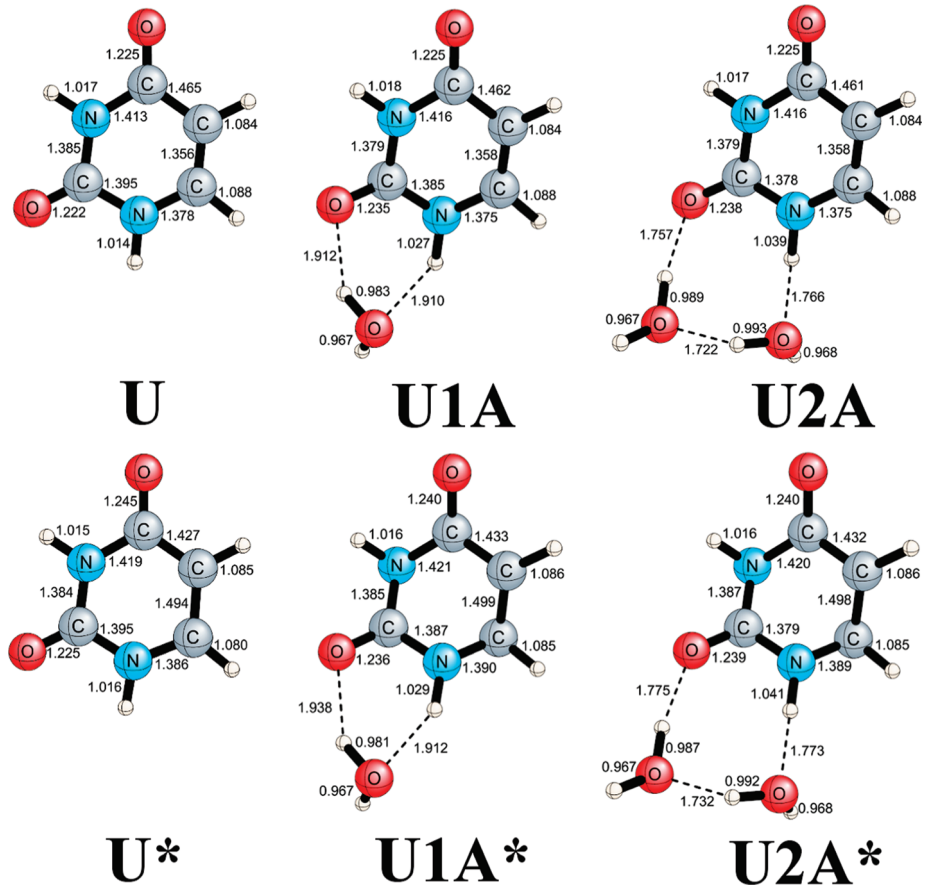


Figure 2. Structures of uracil and its lowest-energy mono- and dihydrates in the ground and lowest triplet states, optimized at the B3LYP/DZP++ level of theory.

defined as the change in absolute energy for a ground state equilibrium geometry upon photonic excitation. The singlet−triplet gap is the difference between the energies of the optimized ground state geometry and the optimized triplet state geometry.

$$\text{VEx} = E(\text{triplet energy at optimized singlet geometry}) - E(\text{optimized singlet})$$

$$\text{singlet−triplet gap} = E(\text{optimized triplet}) - E(\text{optimized singlet})$$

Results

Predicted relative energies of the pyrimidine mono- and dihydrates are listed in Table 1. Figures 2, 3, and 4 display the lowest-energy structures for the free bases and their hydrates in both the ground and the lowest triplet states. All optimized structures have been included as Supporting Information and their intermolecular hydrogen bond lengths are summarized in Tables 2 and 3. The numbering formalism used is as follows: C, T, or U denoting cytosine, thymine or uracil structure, respectively, followed by a 1 or 2 indicating the number of water molecules included, and a capital letter for the relative energetic ordering within each set of either singly or doubly hydrated pyrimidines (i.e., C2A < C2B < C2C). The corresponding triplet structure is indicated by an asterisk (*) and the letter may not indicate the relative energetic ordering of the triplet structures (i.e., in the case of cytosine, C1B* is the lowest-energy triplet structure, not C1A*).

Molecular Structures and Relative Energies. The lowest-energy monohydrate structure of the ground-state uracil (U1A) is characterized by a $N_1-H\cdots O_w-H_w\cdots O_2=C_2$ cyclic hydrogen bond between the uracil base and the water molecule. The $N_1-H\cdots O_w$ and $C_2=O_2\cdots H_w$ hydrogen bond lengths are computed to be 1.910 and 1.912 Å, respectively. The lowest-energy structure of uracil dihydrate in the ground state (U2A) also has a cyclic hydrogen bond that connects the uracil moiety and two water molecules. However, its intermolecular distances are shorter than those of monohydrate U1A, with the $N_1-H\cdots O_w$, $C_2=O_2\cdots H_w$, and $O_w-H_w\cdots O_w$ hydrogen bond lengths being 1.766, 1.757, and 1.722 Å, respectively. The excitation of U1A and U2A to their lowest-triplet states (U1A* and U2A*) increases the intermolecular hydrogen bond distances, implying that stabilization through hydration is reduced in the triplet state, compared to the ground state. Other uracil mono- and dihydrate structures also show a similar weakening of the intermolecular hydrogen bond upon excitation to their lowest-triplet states.

Structural differences between the ground state and the lowest-triplet state of thymine hydrates are similar to those of uracil hydrates in general. In addition, the different hydrated ground state and triplet state thymine and uracil structures display the same energetic ordering. However, in the case of cytosine, neither the singly nor doubly hydrated isomers have the same energetic ordering for the ground state and the triplet state. While the lowest-energy structures of

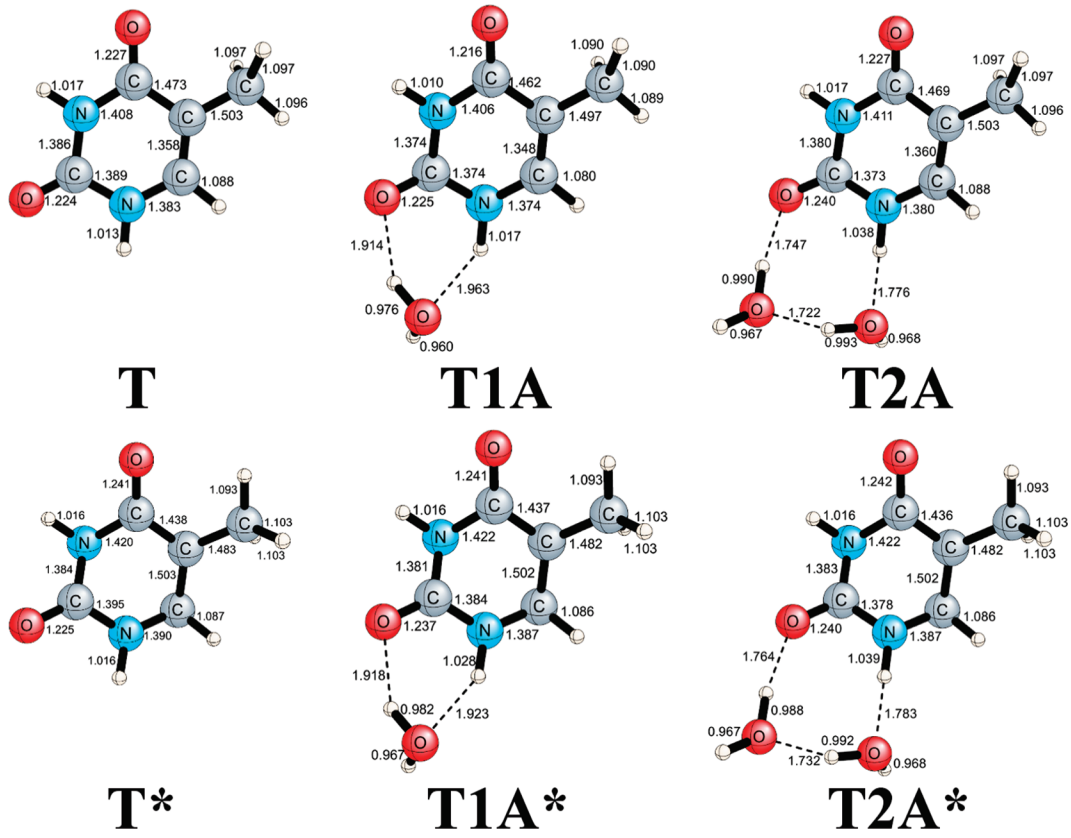


Figure 3. Structures of thymine and its lowest-energy mono- and dihydrates in the ground and lowest triplet states, optimized at the B3LYP/DZP++ level of theory.

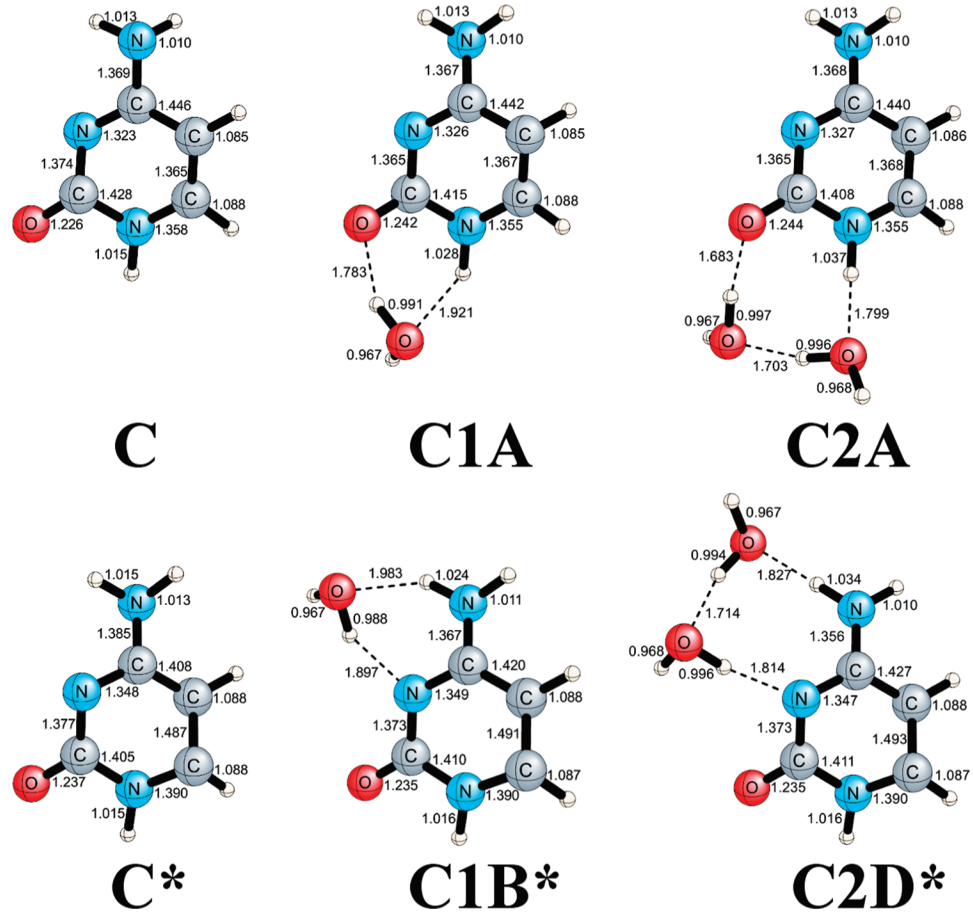


Figure 4. Structures of cytosine and its lowest-energy mono- and dihydrates in the ground and lowest triplet states, optimized at the B3LYP/DZP++ level of theory.

Table 2. Intermolecular Hydrogen Bond Distances in Å for Singly and Doubly Hydrated Uracil and Thymine in the Ground States and the Lowest Triplet Excited States

structure	parameter[a]	singlet	triplet	structure	parameter[a]	singlet	triplet
	uracil monohydrates				thymine monohydrates		
U1A	$N_1-H\cdots O_w$	1.910	1.912	T1A	$N_1-H\cdots O_w$	1.923	1.923
	$C_2=O_2\cdots H_w$	1.912	1.938		$C_2=O_2\cdots H_w$	1.894	1.918
U1B	$N_3-H\cdots O_w$	1.952	1.968	T1B	$N_3-H\cdots O_w$	1.963	1.981
	$C_4=O_4\cdots H_w$	1.891	1.925		$C_4=O_4\cdots H_w$	1.886	1.898
U1C	$C_2=O_2\cdots H_w$	1.937	1.960	T1C	$C_2=O_2\cdots H_w$	1.917	1.939
	$N_3-H\cdots O_w$	1.977	1.989		$N_3-H\cdots O_w$	1.986	2.006
U1D	$C_4=O\cdots H_w$	1.882	1.905	T1D	$C_4=O_4\cdots H_w$	1.878	1.867
	uracil dihydrates				thymine dihydrates		
U2A	$N_1-H\cdots O_{w1}$	1.766	1.773	T2A	$N_1-H\cdots O_{w1}$	1.776	1.783
	$O_{w1}-H_{w1}\cdots O_{w2}$	1.722	1.732		$O_{w1}-H_{w1}\cdots O_{w2}$	1.722	1.732
	$C_2=O_2\cdots H_{w2}$	1.757	1.775		$C_2=O_2\cdots H_{w2}$	1.747	1.764
U2B	$N_3-H\cdots O_{w1}$	1.765	1.777	T2B	$N_3-H\cdots O_{w1}$	1.771	1.786
	$O_{w1}-H_{w1}\cdots O_{w2}$	1.754	1.762		$O_{w1}-H_{w1}\cdots O_{w2}$	1.757	1.764
	$C_4=O_4\cdots H_{w2}$	1.759	1.780		$C_4=O_4\cdots H_{w2}$	1.754	1.764
U2C	$N_1-H\cdots O_{w1}$	1.905	1.912	T2C	$N_1-H\cdots O_{w1}$	1.919	1.923
	$C_2=O_2\cdots H_{w1}$	1.900	1.934		$C_2=O_2\cdots H_{w1}$	1.883	1.914
	$N_3-H\cdots O_{w2}$	1.949	1.975		$N_3-H\ldots O_{w2}$	1.956	1.984
	$C_4=O_4\cdots H_{w2}$	1.884	1.922		$C_4=O_4\cdots H_{w2}$	1.881	1.895
U2D	$C_2=O_2\cdots H_{w1}$	1.785	1.799	T2D	$C_2=O_2\cdots H_{w1}$	1.773	1.788
	$O_{w1}\cdots H_{w2}-O_{w2}$	1.786	1.790		$O_{w1}\cdots H_{w2}-O_{w2}$	1.783	1.791
	$N_3-H\cdots O_{w2}$	1.767	1.779		$N_3-H\cdots O_{w2}$	1.773	1.788
U2E	$N_1-H\cdots O_{w1}$	1.907	1.909	T2E	$N_1-H\cdots O_{w1}$	1.919	1.919
	$C_2=O_2\cdots H_{w1}$	1.920	1.944		$C_2=O_2\cdots H_{w1}$	1.905	1.926
	$C_2=O_2\cdots H_{w2}$	1.950	1.963		$C_2=O_2\cdots H_{w2}$	1.930	1.945
	$N_3-H\cdots O_{w2}$	1.973	1.988		$N_3-H\cdots O_{w2}$	1.980	2.002
U2F	$N_1-H\cdots O_{w1}$	1.901	1.906	T2F	$N_1-H\cdots O_{w1}$	1.909	1.913
	$C_2=O_2\cdots H_{w1}$	1.929	1.957		$C_2=O_2\cdots H_{w1}$	1.917	1.941
	$C_4=O_4\cdots H_{w2}$	1.879	1.906		$C_4=O_4\cdots H_{w2}$	1.878	1.868
U2G	$C_2=O_2\cdots H_{w1}$	1.952	1.978	T2G	$C_2=O_2\cdots H_{w1}$	1.940	1.960
	$N_3-H\cdots O_{w1}$	1.957	1.973		$N_3-H\cdots O_{w1}$	1.957	1.984
	$C_4=O_4\cdots H_{w2}$	1.870	1.898		$C_4=O_4\cdots H_{w2}$	1.871	1.864
U2H	$N_3-H\cdots O_{w1}$	1.935	1.954	T2H	$N_3-H\cdots O_{w1}$	1.939	1.956
	$C_4=O_4\cdots H_{w1}$	1.928	1.954		$C_4=O_4\cdots H_{w1}$	1.927	1.939
	$C_4=O_4\cdots H_{w2}$	1.886	1.905		$C_4=O_4\cdots H_{w2}$	1.891	1.868

[a] Oxygen and hydrogen atoms of a water molecule are denoted with O_w and H_w, respectively. Subscripts w1 and w2 are used to distinguish two different water molecules in the dihydrate structures.

the ground-state cytosine mono- and dihydrates (C1A and C2A) have a $N_1-H\cdots(O_w-H_w)_n\cdots O_2=C_2$ cyclic hydrogen bond, the water molecules in the lowest-energy structures of the triplet-state cytosine hydrates (C1B* and C2D*) bind to the cytosine unit through the N_3 and N_4-H atoms of cytosine, implying that the intermolecular hydrogen bonding in C1B* and C2D* stabilizes the system more strongly than C1A* and C2A*, respectively.

In general, the spread in relative energies for the triplet state structures is approximately 0.5 kcal mol^{-1} smaller than that of their ground state analogs. The SCIPCM method predicted that the effects of macrosolvation narrow the energy differences among the different hydrate structures for the three pyrimidines, but do not significantly change the relative orderings for all three pyrimidine hydrates both in the ground and in the lowest triplet states. Especially for all the nucleobases, the lowest-energy mono- and dihydrate structures in the gas phase were also predicted to be energetically favorable in the condensed media.

Excitation Energies. Estimated vertical and adiabatic excitation energies of the bases and their hydrates are reported in Table 4. In general, cytosine and its hydrates show the largest vertical and adiabatic excitation energies, while thymine and its hydrates have the smallest excitation energies. The methyl group of the unhydrated thymine lowers the vertical and adiabatic excitation energies by 0.13 and 0.18 eV, respectively, compared to those of free uracil. In the thymine hydrates, the methyl group also lowers the excitation energies by similar magnitudes (0.11−0.16 eV for and 0.17−0.20 eV for the vertical and adiabatic excitation energies, respectively). A consequence of this methyl group effect is that the excitation energies of uracil and its hydrates become closer to those of cytosine and its hydrates, compared to the thymine hydrates.

The singlet−triplet gaps of the gas phase uracil, thymine, and cytosine are estimated to be 2.92, 2.74, and 2.97 eV, respectively. The addition of one and two water molecules to thymine and uracil causes a negligible increase in the vertical excitation energies and singlet−triplet gap, as shown in Figure 5. For the cytosine molecule, the effect is slightly more pronounced, with the addition of each water molecule adding ∼0.05 to 0.1 eV to the vertical excitation energies and singlet−triplet gap. In addition, the macrohydration effects estimated using the SCIPCM method were not predicted to cause a significant change in both the vertical and adiabatic excitation energies for the three nucleobases and their hydrates. These results are encouraging in suggesting that the hydrated species are well approximated by the isolated pyrimidines.

Table 3. Intermolecular Hydrogen Bond Distances in Å for Singly and Doubly Hydrated Cytosine in the Ground States and the Lowest Triplet Excited States

structure	parameter[a]	singlet	triplet
	cytosine monohydrates		
C1A	$N_1-H\cdots O_w$	1.921	1.965
	$C_2{=}O_2\cdots H_w$	1.783	1.889
C1B	$N_3\cdots H_w-O_w$	1.894	1.897
	$N_4-H\cdots O_w$	1.981	1.983
C1C	$N_4-H\cdots O_w$	2.023	2.051
	cytosine dihydrates		
C2A	$N_1-H\cdots O_{w1}$	1.799	1.828
	$O_{w1}-H_{w1}\cdots O_{w2}$	1.703	1.743
	$C_2{=}O_2\cdots H_{w2}$	1.683	1.762
C2B	$N_3\cdots H_{w1}-O_{w1}$	1.819	1.811
	$O_{w1}\cdots H_{w2}-O_{w2}$	1.731	1.729
	$N_4-H\cdots O_{w2}$	1.841	1.826
C2C	$N_1-H\cdots O_{w1}$	1.917	1.953
	$C_2{=}O_2\cdots H_{w1}$	1.796	1.881
	$N_3\cdots H_{w2}-O_{w2}$	1.901	1.892
	$N_4-H\cdots O_{w2}$	1.980	1.970
C2D[b]	$C_2{=}O_2\cdots H_{w1}$	1.904	
	$N_3\cdots H_{w1}-O_{w1}$		1.814
	$O_{w1}\cdots H_{w2}-O_{w2}$	1.857	1.714
	$N_4-H\cdots O_{w2}$	1.892	1.827
C2E	$N_1-H\cdots O_{w1}$	1.945	1.989
	$C_2{=}O_2\cdots H_{w1}$	1.758	1.846
	$N_4-H\cdots O_{w2}$	2.018	2.031
C2F	$N_3\cdots H_w-O_{w1}$	1.873	1.868
	$N_4-H\cdots O_{w1}$	2.037	2.025
	$N_4-H\cdots O_{w2}$	2.027	2.029

[a] Oxygen and hydrogen atoms of a water molecule are denoted with O_w and H_w, respectively. Subscripts w1 and w2 are used to distinguish two different water molecules in the dihydrate structures. [b] Upon excitation of C2D to its lowest triplet excited states, the $C_2{=}O_2\cdots H_{w1}$ bond breaks and the $N_3\cdots H_{w1}-O_{w1}$ forms.

Hydration Energies and Dipole Moments. Table 5 reports hydration energies of the different hydrate structures of the three bases in the ground and lowest triplet states. For comparison purposes, those of the anionic hydrates, computed at the same level of theory, are also included. In all singly and doubly hydrated structures of cytosine, thymine, and uracil considered, with the exception of C2D*, the hydration energy of the triplet state is less (~0.5 to 3.5 kcal mol^{-1}) than that of the corresponding singlet structure. The difference is more pronounced for the cytosine structures than for thymine and uracil. The decrease in the hydration energy upon the excitation from the ground state to the lowest-triplet state for all three bases is attributed to the smaller dipole moment in the lowest triplet states, as shown in Table 6, which compares the dipole moments of the three bases and their hydrates in the ground and lowest triplet states. In a recent study on 4-thiouracil by Shukla and Leszczynski,[84] a similar decrease in dipole moment was predicted upon the excitation from the ground to the lowest singlet excited state.

For the cytosine hydrates, the hydration energy change upon the excitation of the hydrated structures may alter their relative energy ordering compared to that of the corresponding neutrals. For example, the excitation from C2A to C2A* is accompanied by a large decrease in the hydration energy by 3.6 kcal mol^{-1} (from 19.9 to 16.3 kcal mol^{-1}). On the contrary, the excitation from C2D to C2D* increases the hydration energy by 0.4 kcal mol^{-1} (from 17.5 kcal mol^{-1} to 17.9 kcal mol^{-1}). As a consequence, the energy ordering between C2A* and C2D* is swapped, compared to that between the corresponding neutrals, C2A and C2D. Therefore, it is the change in hydration energy upon excitation that alters the relative energetic ordering of the lowest-triplet hydrate structures of cytosine. For the uracil and thymine hydrates, the changes in hydration energy upon excitation do not vary enough to affect their energetic orderings.

Table 4. Vertical and Adiabatic Excitation Energies in eV for Singly and Doubly Hydrated Uracil, Thymine, And Cytosine[a]

	VE_x		singlet−triplet gap		
structure	gas	SCIPCM	gas		SCIPCM
	uracil and its hydrates				
U	3.56	3.59	3.04	(2.92)	3.09
U1A	3.57	3.60	3.07	(2.95)	3.11
U1B	3.56	3.59	3.06	(2.95)	3.11
U1C	3.59	3.62	3.06	(2.94)	3.11
U1D	3.54	3.58	3.05	(2.93)	3.11
U2A	3.58	3.61	3.09	(2.97)	3.12
U2B	3.56	3.60	3.07	(2.96)	3.15
U2C	3.57	3.60	3.11	(2.99)	3.14
U2D	3.60	3.63	3.06	(2.94)	3.11
U2E	3.60	3.63	3.09	(2.97)	3.12
U2F	3.55	3.58	3.08	(2.96)	3.12
U2G	3.58	3.61	3.07	(2.95)	3.12
U2H	3.53	3.57	3.07	(2.96)	3.13
	thymine and its hydrates				
T	3.43	3.45	2.85	(2.74)	2.89
T1A	3.44	3.46	2.88	(2.76)	2.90
T1B	3.42	3.44	2.87	(2.75)	2.90
T1C	3.47	3.49	2.87	(2.76)	2.91
T1D	3.39	3.42	2.85	(2.74)	2.90
T2A	3.45	3.47	2.90	(2.78)	2.92
T2B	3.42	3.45	2.88	(2.77)	2.91
T2C	3.43	3.45	2.91	(2.79)	2.92
T2D	3.49	3.50	2.88	(2.77)	2.92
T2E	3.48	3.49	2.90	(2.78)	2.92
T2F	3.40	3.43	2.89	(2.77)	2.92
T2G	3.44	3.46	2.88	(2.76)	2.92
T2H	3.37	3.40	2.87	(2.76)	2.91
	cytosine and its hydrates				
C	3.54	3.62	3.10	(2.97)	3.19
C1A	3.63	3.68	3.21	(3.08)	3.27
C1B	3.56	3.62	3.13	(3.00)	3.19
C1C	3.56	3.64	3.11	(2.98)	3.19
C2A	3.67	3.71	3.26	(3.13)	3.30
C2B	3.58	3.62	3.14	(3.01)	3.18
C2C	3.64	3.66	3.22	(3.10)	3.26
C2D	3.65	3.69	3.06	(2.95)	3.09
C2E	3.64	3.69	3.22	(3.08)	3.26
C2F	3.56	3.62	3.12	(3.00)	3.18

[a] Zero-point vibrational energy (ZPVE)—corrected values are in parentheses.

Discussion

In order to better understand hydration effects upon excitation of the three pyrimidine bases from the ground state to the lowest triplet state, it would be helpful to consider two pathways through which the ground-state unhydrated base changes into the hydrate of the triplet state. As displayed in Figure 6(a), one involves the excitation from the ground state to the triplet state of the unhydrated base, followed by

Table 5. Comparison of Hydration Energies (kcal mol^{-1}) of the Neutral Hydrates of Uracil, Thymine, And Cytosine in the Ground and the Triplet Excited States with the Corresponding Anionic Hydrates Computed at the B3LYP/DZP++ Level of Theory[a]

	neutral					
structure	ground state		triplet state		anion[b]	
			uracil hydrates			
U1A	10.7	(8.5)	9.9	(7.8)	13.1	(10.9)
U1B	9.1	(6.9)	8.5	(6.4)	14.8	(12.9)
U1C	8.4	(6.3)	8.0	(6.0)	12.0	(10.2)
U1D	7.3	(5.4)	7.1	(5.3)	15.0	(13.2)
U2A	23.4	(18.5)	22.1	(17.3)	24.4	(19.9)
U2B	20.8	(16.1)	20.0	(15.3)	27.3	(22.6)
U2C	20.0	(15.6)	18.4	(14.1)	27.6	(23.4)
U2D	19.4	(15.0)	18.9	(14.4)	27.2	(22.6)
U2E	18.6	(14.4)	17.5	(13.4)	23.9	(19.8)
U2F	17.9	(13.8)	16.9	(12.8)	27.6	(23.6)
U2G	15.9	(12.0)	15.2	(11.3)	26.3	(22.6)
U2H	16.0	(12.0)	15.3	(11.2)	28.6	(24.7)
			thymine hydrates			
T1A	10.7	(8.4)	10.0	(7.8)	13.0	(10.9)
T1B	9.0	(6.8)	8.5	(6.4)	14.7	(12.7)
T1C	8.6	(6.5)	8.0	(6.0)	11.9	(10.1)
T1D	6.4	(4.6)	6.3	(4.6)	14.2	(12.4)
T2A	23.3	(18.5)	22.1	(17.4)	24.4	(19.8)
T2B	20.5	(15.9)	19.9	(15.2)	27.1	(22.3)
T2C	19.8	(15.5)	18.5	(14.3)	27.4	(23.2)
T2D	19.5	(15.1)	18.8	(14.4)	27.1	(22.3)
T2E	18.8	(14.5)	17.6	(13.4)	23.8	(19.6)
T2F	17.1	(13.1)	16.2	(12.4)	26.8	(22.8)
T2G	15.2	(11.4)	14.4	(10.7)	25.5	(21.8)
T2H	14.8	(10.9)	14.4	(10.5)	27.5	(23.7)
			cytosine hydrates			
C1A	12.1	(9.7)	9.4	(7.1)	15.7	(13.7)
C1B	11.5	(9.1)	10.7	(8.4)	17.7	(15.3)
C1C	6.0	(4.5)	5.8	(4.3)	9.7	(6.6)
C2A	24.8	(19.9)	21.1	(16.3)	27.6	(23.7)
C2B	23.5	(18.6)	22.6	(17.7)	31.4	(26.8)
C2C	23.3	(18.6)	20.4	(15.8)	32.3	(27.9)
C2D	22.0	(17.5)	22.9	(17.9)	33.6	(28.7)
C2E	18.2	(14.2)	15.4	(11.7)	22.7	(18.4)
C2F	17.5	(13.6)	17.0	(13.0)	22.2	(18.0)

[a] Zero-point vibrational energy (ZPVE)—corrected results are in parentheses. [b] Refs., 73, 76 and 58 for uracil, thymine, and cytosine, respectively.

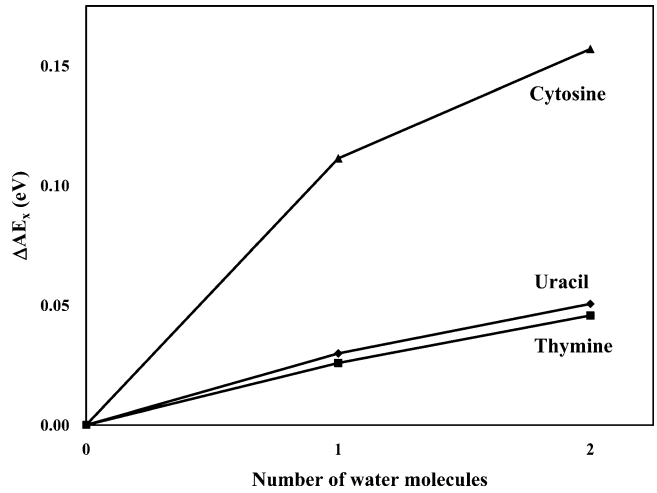


Figure 5. Changes in adiabatic excitation energies (eV) for the lowest energy structure upon addition of one and two water molecules for the three pyrimidines.

Table 6. Dipole Moments (μ, Debye) of the Neutral Hydrates of Uracil, Thymine, And Cytosine in the Ground and the Triplet Excited States, Computed at the B3LYP/DZP++ Level of Theory

structure	μ	structure	μ
	uracil hydrate		
U	4.63	U*	3.91
U1A	4.00	U1A*	3.31
U1B	4.57	U1B*	3.95
U1C	5.16	U1C*	4.31
U1D	2.95	U1D*	2.25
U2A	3.72	U2A*	2.98
U2B	4.14	U2B*	3.49
U2C	3.90	U2C*	3.05
U2D	5.12	U2D*	4.57
U2E	5.16	U2E*	4.40
U2F	2.97	U2F*	2.25
U2G	3.78	U2G*	2.98
U2H	2.57	U2H*	2.12
	thymine hydrate		
T	4.59	T*	4.28
T1A	3.69	T1A*	3.29
T1B	4.79	T1B*	4.69
T1C	5.01	T1C*	4.44
T1D	3.68	T1D*	3.28
T2A	3.36	T2A*	2.97
T2B	4.34	T2B*	4.18
T2C	3.93	T2C*	3.57
T2D	4.96	T2D*	4.77
T2E	4.74	T2E*	4.14
T2F	3.74	T2F*	3.37
T2G	4.64	T2G*	4.09
T2H	3.31	T2H*	3.17
	cytosine hydrate		
C	6.79	C*	5.28
C1A	5.71	C1A*	4.49
C1B	6.23	C1B*	5.10
C1C	9.64	C1C*	8.41
C2A	5.23	C2A*	3.98
C2B	5.99	C2B*	5.19
C2C	4.84	C2C*	3.84
C2D	5.42	C2D*	4.97
C2E	9.01	C2E*	8.12
C2F	8.73	C2F*	8.05

subsequent hydration. In the other process, formation of the hydrate of the ground-state base occurs first, and then the hydrate is excited to the triplet state. Because the energy difference between the initial state (the unhydrated base at its ground state equilibrium geometry) and the final state (the hydrate of the triplet-state base) should be independent of the two pathways, an equation for hydration effects on the excitation energy of a DNA/RNA base can be derived:

$$E_{ex}(B) - E_{hyd}(B^*) = E_{ex}(B \cdot H_2O) - E_{hyd}(B)$$

$$E_{ex}(B \cdot H_2O) - E_{ex}(B) = -[E_{hyd}(B^*) - E_{hyd}(B)]$$

$$\Delta E_{ex}(B \cdot H_2O, B) = -\Delta E_{hyd}(B^*, B)$$

where B and B* denote a nucleobase in its ground and triplet-excited states, respectively. The equation above shows that the change in the excitation energy upon hydration, $\Delta E_{ex}(B \cdot H_2O, B)$, is equivalent to the negative value of the difference in hydration energy between the ground and triplet states of the unhydrated base, $\Delta E_{hyd}(B^*,B)$. As shown in Table 5, the hydration energy of the triplet state is smaller

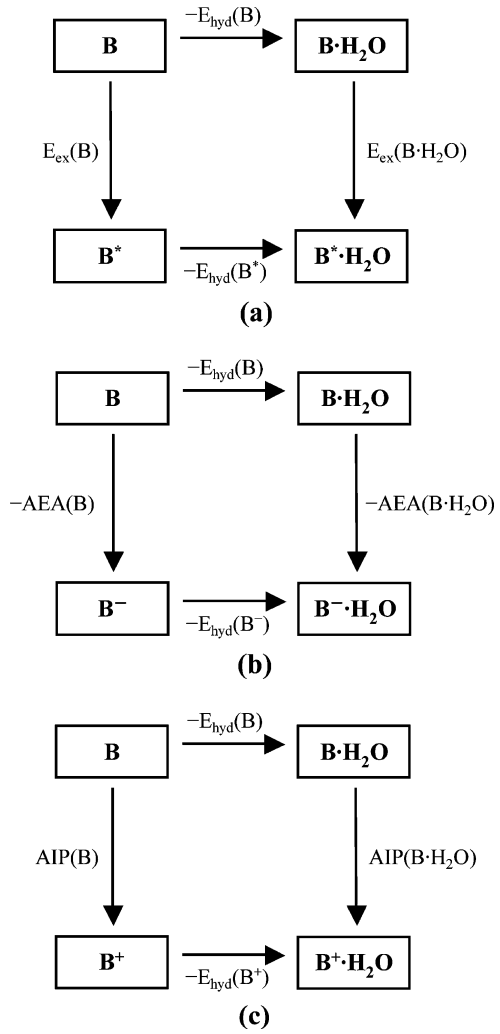


Figure 6. Two potential pathways in which the unhydrated neutral nucleobase in the ground state (B) is converted into (a) the hydrate of the neutral base in the lowest triplet state (B*); (b) the anionic hydrate of the base (B^-); and (c) the cationic hydrate of the base (B^+). E_{ex}, AEA, and AIP represent the excitation energy, adiabatic electron affinity, and adiabatic ionization potential, respectively.

than that of the ground state, making the $\Delta E_{hyd}(B^*,B)$ term negative and hence, the $\Delta E_{ex}(B\cdot H_2O, B)$ term positive. However, note that the magnitude of the $\Delta E_{hyd}(B^*,B)$ term for monohydrates of uracil and thymine, which is only 0.4 kcal mol^{-1} on average, indicates that the change in excitation energy upon hydration, $\Delta E_{ex}(B\cdot H_2O, B)$, is less than 0.02 eV. The corresponding value for the cytosine monohydrates is 1.1 kcal mol^{-1} (~0.05 eV).

Many experimental and theoretical studies have shown that, in spite of the near-zero electron affinities of the gas-phase DNA/RNA bases, hydration with even a single water molecule causes an increase in the adiabatic electron affinities of the three pyrimidine bases (by as much as 0.3 eV) and successive addition of more water molecule increases the AEA values even more.[58,73,76] These effects enable negative charge formation on the DNA bases in aqueous solution, leading to lethal DNA lesions through subsequent single- or double-strand breaks. On the contrary, the excitation energies of the three nucleobases computed in the present study are insensitive to the hydration effects, compared to their electron

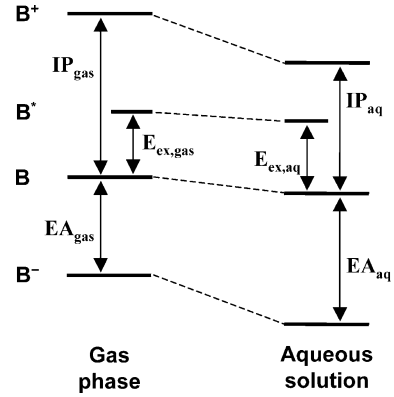


Figure 7. Schematic energy diagram showing the hydration effects upon the excitation energy, electron affinity and ionization potential of the DNA/RNA bases.

affinities. Therefore, it may be helpful to compare anion hydration with the triplet state hydration in the present study. As shown in Figure 6(b), there are also two possible pathways for forming the anionic hydrates of the bases from the unhydrated bases. Similar to the triplet state cases, the changes in electron affinities of the bases upon hydration can be correlated with the changes in the hydration energy between anion (B^-) and neutral bases (B).

$$\Delta AEA(B\cdot H_2O,B) = \Delta E_{hyd}(B^-, B)$$

That is, the change in adiabatic electron affinity, $\Delta AEA(B\cdot H_2O,B)$, is equivalent to the difference between the base and its anion, $\Delta E_{hyd}(B^-, B)$. As shown in Table 5, the hydration energies of the anions, essentially greater than those of neutrals, are responsible for the increase in the AEA upon hydration. In a similar manner, Figure 6(c) can be used to show that the hydration effect on the ionization potential is equivalent to the difference in the hydration energy between the cation (B^+) and the neutral (B).

$$\Delta AIP(B\cdot H_2O,B) = -\Delta E_{hyd}(B^+, B)$$

Although no studies have been reported on the hydration energies of the cations that are computed at the same level of theory employed in the present study, many experimental evidence[85] as well as other theoretical studies[86,87] using different levels of theory showed a significant increase in hydration energy for the cation, compared to the corresponding neutrals, leading to a significant decrease in the ionization potentials of the DNA/RNA bases. For example, while monohydration of thymine is predicted in the present study to increase its excitation energy by only 0.02 eV, its ionization potential upon monohydration was computed to decrease by 0.1 eV (at the B3LYP/6-31+G** level of study)[86,87] and the experimentally determined decrease (by 0.3 eV) is even more significant.[85] Figure 7 displays a schematic energy diagram that compares hydration effects upon the electron affinities, ionization potentials, and excitation energies of the nucleobases. Stabilization due to hydration is more significant in charged species than in neutral species, causing the increase in the electron affinity and the decrease in the ionization potential. On the other hand,

because both the ground state and the excited state of the nucleobases are neutral, the hydration energy difference between the two states is relatively small, compared to that between the neutral and charged bases, making the excitation energies of the nucleobases less sensitive to hydration than their electron affinities and ionization potentials.

Conclusions

The singlet ground states and lowest triplet states of mono- and dihydrates of the three DNA/RNA pyrimidine bases, cytosine, uracil, and thymine, have been investigated at the B3LYP/DZP++ level of theory. For uracil and thymine, the energetic ordering of hydrate structures of the triplet states is the same as that of the corresponding singlet ground states. For all three bases, it was found that hydration does not have a significant effect upon the energy difference between the singlet and triplet states, compared to hydration effects on electron affinities and ionization potentials, which involve charged species. A water molecule is likely to interact with a charged species more strongly than with a neutral species, resulting in the increase in electron affinity and the decrease in ionization potential. On the contrary, if a molecule is neutral in the ground state, its triplet state is necessarily also neutral, and as shown in the present study, the differential stabilization due to hydration of the triplet state is quite modest, making the excitation energy relatively insensitive to hydration.

Acknowledgment. This work was supported by National Science Foundation Grant CHE-0749868. Figures were generated using the software cheMVP, developed by Dr. Andrew Simmonett.

Supporting Information Available: Cartesian coordinates, absolute energies, and vibrational frequencies for the optimized structures of the ground and the lowest triplet state of the mono- and dihydrates of the three pyrimidine hydrates. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Pfeifer, G. P.; You, Y. H.; Besaratinia, A. *Mutat. Res.* **2005**, *571*, 19.

(2) Cadet, J.; Sage, E.; Douki, T. *Mutat. Res.* **2005**, *571*, 3.

(3) Beukers, R.; Berends, W. *Biochim. Biophys. Acta* **1960**, *41*, 550.

(4) Mouret, S.; Baudouin, C.; Charveron, M.; Favier, A.; Cadet, J.; Douki, T. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 13765.

(5) Douki, T.; Reynaud-Angelin, A.; Cadet, J.; Sage, E. *Biochemistry* **2003**, *42*, 9221.

(6) Rochette, P. J.; Therrien, J. P.; Drouin, R.; Perdiz, D.; Bastien, N.; Drobetsky, E. A.; Sage, E. *Nucleic Acids Res.* **2003**, *31*, 2786.

(7) Kappes, U. P.; Luo, D.; Potter, M.; Schulmeister, K.; Runger, T. M. *J. Invest. Dermatol.* **2006**, *126*, 667.

(8) You, Y. H.; Lee, D. H.; Yoon, J. H.; Nakajima, S.; Yasui, A.; Pfeifer, G. P. *J. Biol. Chem.* **2001**, *276*, 44688.

(9) Smith, C. A.; Baeten, J.; Taylor, J. S. *J. Biol. Chem.* **1998**, *273*, 21933.

(10) Johnson, R. E.; Prakash, S.; Prakash, L. *Science* **1999**, *283*, 1001.

(11) Sancar, A. *Biochemistry* **1994**, *33*, 2.

(12) Carell, T. *Angew. Chem., Int. Ed.* **1995**, *34*, 2491.

(13) Sancar, A. *Science* **1996**, *272*, 48.

(14) Sancar, A. *Chem. Rev.* **2003**, *103*, 2203.

(15) Kao, Y. T.; Saxena, C.; Wang, L. J.; Sancar, A.; Zhong, D. P. *Cell Biochem. Biophys.* **2007**, *48*, 32.

(16) Weber, S. *Biochim. Biophys. Acta Bioenerg.* **2005**, *1707*, 1.

(17) Wood, R. D.; Mitchell, M.; Sgouros, J.; Lindahl, T. *Science* **2001**, *291*, 1284.

(18) Li, Y. F.; Kim, S. T.; Sancar, A. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 4389.

(19) Ley, R. D. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 4337.

(20) Sage, E. *Photochem. Photobiol.* **1993**, *57*, 163.

(21) Becker, M. M.; Wang, Z. *J. Mol. Biol.* **1989**, *210*, 429.

(22) Mantione, M.; Pullman, B. *Biochim. Biophys. Acta* **1964**, *91*, 387.

(23) Schreier, W. J.; Schrader, T. E.; Koller, F. O.; Gilch, P.; Crespo-Hernandez, C. E.; Swaminathan, V. N.; Carell, T.; Zinth, W.; Kohler, B. *Science* **2007**, *315*, 625.

(24) Zhang, R. B.; Eriksson, L. A. *J. Phys. Chem. B* **2006**, *110*, 7556.

(25) Abouaf, R.; Pommier, J.; Dunet, H.; Quan, P.; Nam, P. C.; Nguyen, M. T. *J. Chem. Phys.* **2004**, *121*, 11668.

(26) Hou, X. J.; Nguyen, M. T. *Chem. Phys.* **2005**, *310*, 1.

(27) Marian, C. M.; Schneider, F.; Kleinschmidt, M.; Tatchen, J. *Eur. Phys. J. D* **2002**, *20*, 357.

(28) Abouaf, R.; Pommier, J.; Dunet, H. *Chem. Phys. Lett.* **2003**, *381*, 486.

(29) Blancafort, L.; Cohen, B.; Hare, P. M.; Kohler, B.; Robb, M. A. *J. Phys. Chem. A* **2005**, *109*, 4431.

(30) Markovitsi, D.; Onidas, D.; Gustavsson, T.; Talbot, F.; Lazzarotto, E. *J. Am. Chem. Soc.* **2005**, *127*, 17130.

(31) Nguyen, M. T.; Zhang, R.; Nam, P. C.; Ceulemans, A. *J. Phys. Chem. A* **2004**, *108*, 6554.

(32) Bosca, F.; Lhiaubet-Vallet, V.; Cuquerella, M. C.; Castell, J. V.; Miranda, M. A. *J. Am. Chem. Soc.* **2006**, *128*, 6318.

(33) Zhang, R. B.; Zeegers-Huyskens, T.; Ceulemans, A.; Nguyen, M. T. *Chem. Phys.* **2005**, *316*, 35.

(34) Scordamaglia, R.; Cavallone, F.; Clementi, E. *J. Am. Chem. Soc.* **1977**, *99*, 5545.

(35) Pullman, A.; Perahia, D. *Theor. Chim. Acta* **1978**, *48*, 29.

(36) Pullman, B.; Miertus, S.; Perahia, D. *Theor. Chim. Acta* **1979**, *50*, 317.

(37) Sagarik, K.; Corongiu, G.; Clementi, E. *J. Mol. Struct.* **1991**, *81*, 355.

(38) Broo, A.; Holmen, A. *J. Phys. Chem. A* **1997**, *101*, 3589.

(39) Aleman, C. *Chem. Phys. Lett.* **1999**, *302*, 461.

(40) Shishkin, O.; Gorb, L.; Leszczynski, J. *Int. J. Mol. Sci.* **2000**, *1*, 17.

(41) van Mourik, T.; Benoit, D. M.; Price, S. L.; Clary, D. C. *Phys. Chem. Chem. Phys.* **2000**, *2*, 1281.

(42) Civcir, P. U. *J. Mol. Struct.* **2000**, *532*, 157.

(43) Rejnek, J.; Hanus, M.; Labelac, M.; Ryjacek, F.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2006.

(44) Kabelac, M.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 903.

(45) Sobolewski, A. L.; Adamowicz, L. *J. Chem. Phys.* **1995**, *102*, 5708.

(46) Gorb, L.; Leszczynski, J. *Int. J. Quantum Chem.* **1998**, *70*, 855.

(47) Gorb, L.; Podolyan, Y.; Leszczynski, J. *J. Mol. Struct.* **1999**, *487*, 47.

(48) Aleman, C. *Chem. Phys.* **1999**, *244*, 151.

(49) Chandra, A. K.; Nguyen, M. T.; Zeegers-Huyskens, T. *J. Mol. Struct.* **2000**, *519*, 1.

(50) Shishkin, O. V.; Gorb, L.; Leszczynski, J. *J. Phys. Chem. B* **2000**, *104*, 5357.

(51) Sivanesan, D.; Babu, K.; Gadre, S. R.; Subramanian, V.; Ramasami, T. *J. Phys. Chem. A* **2000**, *104*, 10887.

(52) Fogarasi, G.; Szalay, P. G. *Chem. Phys. Lett.* **2002**, *356*, 383.

(53) Trygubenko, S. A.; Bogdan, T. V.; Rueda, M.; Orozco, M.; Luque, F. J.; Sponer, J.; Slavicek, P.; Hobza, P. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4192.

(54) Shukla, M. K.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 11338.

(55) Chandra, A. K.; Michalska, D.; Wysokinsky, R.; Zeegers-Huyskens, T. *J. Phys. Chem. A* **2004**, *108*, 9593.

(56) Hunter, K. C.; Rutledge, L. R.; Wetmore, S. D. *J. Phys. Chem. A* **2005**, *109*, 9554.

(57) Hunter, K. C.; Wetmore, S. D. *Chem. Phys. Lett.* **2006**, *422*, 500.

(58) Kim, S.; Schaefer, H. F. *J. Chem. Phys.* **2007**, *126*, 064301.

(59) Del Bene, J. E. *J. Comput. Chem.* **1981**, *2*, 188.

(60) Del Bene, J. E. *J. Comput. Chem.* **1981**, *2*, 416.

(61) Scheiner, S. *Biopolymers* **1983**, *22*, 731.

(62) Rybak, S.; Szalewicz, K.; Jeziorski, B.; Corongiu, G. *Chem. Phys. Lett.* **1992**, *199*, 567.

(63) Smets, J.; McCarthy, W. J.; Adamowicz, L. *J. Phys. Chem.* **1996**, *100*, 14655.

(64) Ghomi, M.; Aamouche, A.; Cadioli, B.; Berthier, G.; Grajcar, L.; Baron, M. H. *J. Mol. Struct.* **1997**, *410*, 323.

(65) Aamouche, A.; Berthier, G.; Cadioli, B.; Gallinella, E.; Ghomi, M. *J. Mol. Struct.* **1998**, *426*, 307.

(66) van Mourik, T.; Price, S. L.; Clary, D. C. *J. Phys. Chem. A* **1999**, *103*, 1611.

(67) Bencivenni, L.; Ramondo, F.; Pieretti, A.; Sanna, N. *J. Chem. Soc., Perkin Trans. 2* **2000**, 1685.

(68) Gadre, S. R.; Babu, K.; Rendell, A. P. *J. Phys. Chem. A* **2000**, *104*, 8976.

(69) Gaigeot, M. P.; Kadri, C.; Ghomi, M. *J. Mol. Struct.* **2001**, *565*, 469.

(70) Di Laudo, M.; Whittleton, S. R.; Wetmore, S. D. *J. Phys. Chem. A* **2003**, *107*, 10406.

(71) Danilov, V. I.; van Mourik, T.; Poltev, V. I. *Chem. Phys. Lett.* **2006**, *429*, 255.

(72) Bao, X. G.; Sun, H.; Wong, N. B.; Gu, J. D. *J. Phys. Chem. B* **2006**, *110*, 5865.

(73) Kim, S.; Schaefer, H. F. *J. Chem. Phys.* **2006**, *125*, 144305.

(74) Del Bene, J. E. *J. Chem. Phys.* **1982**, *76*, 1058.

(75) Chandra, A. K.; Nguyen, M. T.; Zeegers-Huyskens, T. *J. Phys. Chem. A* **1998**, *102*, 6010.

(76) Kim, S.; Wheeler, S. E.; Schaefer, H. F. *J. Chem. Phys.* **2006**, *124*, 204310.

(77) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. *Gaussian 94, Revision E.2*, Gaussian, Inc.: Pittsburgh PA, 1995.

(78) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.

(79) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823.

(80) Lee, T. J.; Schaefer, H. F. *J. Chem. Phys.* **1985**, *83*, 1784.

(81) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(83) Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. *J. Phys. Chem.* **1996**, *100*, 16098.

(84) Shukla, M. K.; Leszczynski, J. *J. Mol. Struct.* **2006**, *771*, 149.

(85) Kim, S. K.; Lee, W.; Herschbach, D. R. *J. Phys. Chem.* **1996**, *100*, 7933.

(86) Close, D. M.; Crespo-Hernandez, C. E.; Gorb, L.; Leszczynski, J. *J. Phys. Chem. A* **2005**, *109*, 9279.

(87) Close, D. M.; Crespo-Hernandez, C. E.; Gorb, L.; Leszczynski, J. *J. Phys. Chem. A* **2006**, *110*, 7485.

CT900478C

JCTC Journal of Chemical Theory and Computation

Multiple Low-Lying States for Compound I of $P450_{cam}$ and Chloroperoxidase Revealed from Multireference Ab Initio QM/MM Calculations

Hui Chen,*,† Jinshuai Song,†,‡ Wenzhen Lai,† Wei Wu,‡ and Sason Shaik*,†

Institute of Chemistry and the Lise Meitner-Minerva Center for Computational Quantum Chemistry, Hebrew University of Jerusalem, Givat Ram Campus, 91904 Jerusalem, Israel, State Key Laboratory of Physical Chemistry and Chemical Engineering, Xiamen University, 361005 Xiamen, P. R. China

Received November 23, 2009

Abstract: The hybrid CASPT2/MM approach is employed to systematically study the ground and low-lying excited states of the ultimate active species of the enzymes $P450_{cam}$ and chloroperoxidase (CPO): the oxoiron(IV)−porphyrin cation−radical $Por^{\bullet+}Fe^{IV}$═O(Cys) species, the so-called Compound I (Cpd I). The results underscore the fact that the B3LYP/MM method is quite accurate on the most part. However, the CASPT2/MM energies for the ferryl-pentaradicaloid quartet state and the perferryl $Fe^{V}O$ doublet and quartet states are significantly lower than the B3LYP/MM results. Thus, while the present CASPT2/MM may still overestimate the stability of these states, nevertheless, taken at its face value, the result raises the question whether these states actually contribute to the reactivity of Cpd I. Our paper tries to grapple with this question in view of (a) the recent speculations that the perferryl $Fe^{V}O$ states may be involved in unusual reactivities of Cpd I species (Pan, Z. Z.; Wang, Q.; Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2009**, *131*, 2621−2628) and (b) the DFT/MM results which show that the pentaradicaloid states have intrinsically low barriers for H-abstraction (Altun, A.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2007**, *129*, 8978−8987). The application of CASPT2/MM to high valent transition metal states like the perferryl are far from being trivial, and the experience and insight gained in this study are expected to be helpful for future successful application of this type of method to resolve key issues in P450 reactivity.

Introduction

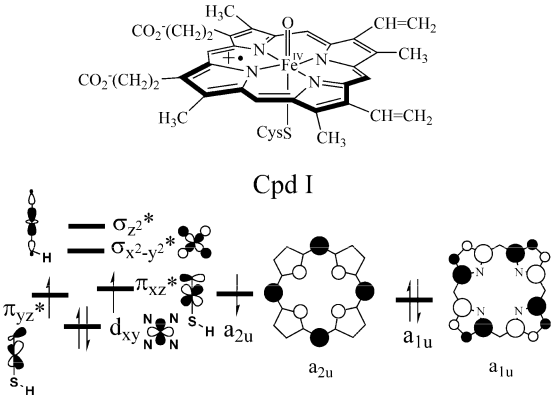
It is generally accepted that the reactive species in the thiolate-ligated heme enzyme cytochromes P450 (P450s)[1,2a] and chloroperoxidase[2] (CPO) is the oxoiron(IV) porphyrin π-cation radical active species $Por^{\bullet+}Fe^{IV}$═O(Cys), termed Compound I (Cpd I), which is responsible for the potent catalytic monooxygenase and peroxidase activities of these enzymes and has the electronic structure shown in Scheme 1. While Cpd I in CPO has been characterized by many spectroscopic methods,[3] the corresponding species in P450s is still elusive in the native catalytic cycle.[1,4] To bypass these difficulties, P450 Cpd I species were generated directly, by reacting the enzyme with *m*-chloroperoxybenzoic acid, and were followed using rapid scan stopped-flow absorption spectroscopy.[5] Very recently, P450 Cpd I species were produced also by laser flash photolysis (LFP) of the one-electron reduced species, in a manner that enabled the following of their reactivity patterns toward a variety of substrates.[6] These studies questioned the consensus that the reactive state for P450 in the native cycle is the π-cation radical ferryl state, $Por^{\bullet+}Fe^{IV}$═O(Cys).

Thus, the elusive nature of Cpd I in the native P450 cycle has generated alternative hypotheses regarding the nature of the "active oxidant" of P450. One of the species that has

* Corresponding author fax: +972 (0)2 658 4033; e-mail: chenh@yfaat.ch.huji.ac.il (H.C.), sason@yfaat.ch.huji.ac.il (S.S.).
† Hebrew University of Jerusalem.
‡ Xiamen University.

10.1021/ct9006234
Published on Web 02/19/2010

Scheme 1. Orbital Diagram for the Doublet Ground State of Cpd I Species in $P450_{cam}$ and CPO on the Basis of DFT/MM Calculation



recently been invoked to explain the high reactivity of $P450_{cam}$ compared with those of the LFP-generated Cpd I species is the perferryl, $PorFe^{V}=O$,[6] in which the porphyrin gets one electron from one of the orbitals (d_{xy} or π^*) in the d-block of iron to become closed-shell (see Scheme 1). Another state that has been revealed by density functional theory (DFT)[7] and DFT/MM calculations[8,9] is the pentaradicaloid state, in which an electron from the d_{xy} (also called δ) orbital of the triradicaloid state, in Scheme 1, is promoted to the $\sigma^*_{x^2-y^2}$ orbital. This latter state has just recently been speculated as a possible cause of the extremely high reactivity of a model Cpd I species.[10]

Due to the prominent role played by Cpd I in P450s chemistry and its elusive status, within the native catalytic cycle, it has been extensively studied by the hybrid quantum mechanical/molecular mechanical (QM/MM) approach,[11] which accounts for effects associated with the protein environment. The major QM tool has so far been DFT/MM, which basically gave results compatible with DFT-only calculations that included bulk polarity and hydrogen bonding (H-bonding) effects due to NH···S interactions with the thiolate ligand. However, it has been obvious that further studies should ultimately be carried out at the ab initio level with the multireference configuration interaction (MRCI) or multireference second-order perturbation (MRPT2) with an adequate basis set and a large active space. In this work, we report such a study of this key species, for the two heme enzymes, based on high level ab initio multiconfigurational second-order perturbation[12] CASPT2/MM calculations, with an aim of establishing the state ordering of all the low-lying states within around 30 kcal/mol, the lower ones of which could possibly contribute to the reactivity of this species.

Three ab initio multireference studies have been reported so far, which addressed either the gas phase species or were limited to a few states.[8,13,14] Thus, Thiel and co-workers used the difference dedicated configuration interaction (DDCI2) method[8] and, later, DDCI2-Q with MR-Davidson correction (which treats the size-consistency problem of CI) to conduct QM/MM calculation on Cpd I of $P450_{cam}$.[13] The quartet a_{2u} triradicaloid state, having the electronic structure in Scheme 1 ($a_{2u}^1\pi^*_{yz}{}^1\pi^*_{xz}{}^1$), but with all electrons being spin-up, was found to be the lowest quartet state in DFT/MM calculations[8,9] but only slightly lower (1.9 kcal/mol) than the corresponding a_{1u} triradicaloid state ($a_{1u}^1\pi^*_{yz}{}^1\pi^*_{xz}{}^1$) at the DDCI2-Q/MM level.[13] However, the corresponding calculation for the doublet state with the large active space encountered convergence problems. So with large active space, no states other than the quartet triradicaloid a_{1u} and a_{2u} states were explored to date using in-protein calculations. Radoń et al.[14] used the CASPT2 method to study a Cpd I model of P450 in the gas phase and found that multistate CASPT2 was crucial for generating the porphyrin-based cation radical state as the ground state. Inclusion of the axial thiolate ligand orbital into the active space seemed to be necessary because the sulfur-based radical state was significantly lower in energy than the porphyrin-based triradicaloid state at the complete active space self-consistent field (CASSCF) level that precedes the CASPT2 calculations. So the situations in gas phase and in protein calculations of Cpd I appear very different. This difference was also exhibited by gas phase DFT studies[15] which revealed a ground state with a high sulfur radical character, vis-à-vis in protein DFT/MM calculations,[8,11] which showed a dominant porphyrin-based radical structure. By contrast to the gas phase results, in CPO, where Cpd I is observable, electron−nuclear double resonance (ENDOR) measurements have shown that the thiolate spin density of Cpd I is less than 0.23,[3e] which supports the DFT/MM[8,11] and the DDCI2/MM results.[13] It is apparent therefore that QM/MM treatments based on ab initio multireference correlated methods for the important open-shell Cpd I system are essential but have still not been used to describe a variety of low lying states that might be contributing to the reactivity of this species. There are many important and interesting questions to be answered in this field. As we mentioned above, DFT[16] and DFT/MM[9] calculations have shown that the pentaradicaloid states of Cpd I in P450 could be more reactive than the triradicaloid ground state. But how high is the pentaradicaloid state in energy relative to the triradicaloid ground state for Cpd I species? B3LYP[7,16] and B3LYP/MM calculations[8,9] for $P450_{cam}$ suggested that this state is adiabatically about 14 and 12 kcal/mol higher than the ground state, respectively. However, previous spectroscopy-oriented configuration interaction (SORCI) calculations on model nonheme iron-oxo systems[8,17] suggested that the state having an iron $d_{xy}{}^1\sigma^*_{x^2-y^2}{}^1$ configuration (two iron d-type orbitals spanning the equatorial plane), analogous to the pentaradicaloid state in Cpd I, and the state having a $d_{xy}{}^2\sigma^*_{x^2-y^2}{}^0$ configuration, analogous to the triradicaloid state in Cpd I, are close in energy. Another intriguing question is how high the perferryl Fe^{V}-oxo species are in Cpd I, which has been repeatedly proposed to be the actual reactive species in native P450 chemistry.[6,18] According to recent TD-DFT/MM calculations of $P450_{cam}$ Cpd I, these Fe^{V} states are at least 37 kcal/mol higher in energy relative to the doublet ground state triradicaloid state (Scheme 1) and hence cannot possibly contribute to reactivity.[9] In order to answer these questions using a high-level wave function theory, we decided to carry out systematic CASPT2/MM calculations of both the ground state of Cpd I and its low lying excited states. Thus, while the present results may certainly not be the last word on Cpd I, still they constitute state-of-the-art values that will provide a challenge to proceed to even higher

QM/MM levels, with an aim of establishing a broader and deeper understanding of the electronic structures of Cpd I of thiolate enzymes and its potential as a multistate reactivity (MSR) reagent.

The Computational Methodology. *DFT/MM Calculations.* The QM/MM setup procedure used here for $P450_{cam}$ and CPO was described extensively in our previous works.[19] Here, we addressed only the essential features and relegated the rest of the details to the Supporting Information (SI) document.

All DFT/MM computations were performed using ChemShell[20] interfaced with Turbomole[21] and DL_POLY.[22] The hybrid B3LYP[23] functional was used for the QM region, and the CHARMM22[24] force field was used for the MM region. The geometries were optimized with the double-ζ LACVP[25] basis set (B1), followed by a single-point energy correction with a larger basis set B2, and the Wachters' all electron basis set[26] augmented with diffuse d and polarization f functions on iron (8s7p4d1f) and 6-31++G(d,p)[27] on the other atoms. The QM region in our QM/MM calculations involves porphine with an iron-oxo unit and axial cysteine ligand modeled as SH. The electronic embedding scheme[28] was used to account for the polarization effect of the QM part induced by the protein environment. No cutoffs were introduced for the nonbonding MM and QM/MM interactions. The dangling bond at the QM/MM boundary was saturated by a hydrogen-link atom and treated in the framework of the charge-shift method.[28] Full geometry optimizations were performed with HDLC optimizer.[29] No symmetry constraints were imposed on the studied Cpd I systems.

CASPT2/CASSCF/MM Calculations. The multireference ab initio CASPT2/CASSCF/MM calculations were carried out with the MOLCAS 7.2 suite of programs[30] using geometries optimized at the B3LYP/MM level. The QM-polarizing point charges generated in the DFT/MM geometry optimization were used during the CASPT2/CASSCF/MM procedure. The Douglas−Kroll−Hess Hamiltonian[31] was used to account for the scalar relativistic effects. A large basis set was required in the CASPT2/CASSCF calculations: On iron, we used a triple-ζ cc-pwCVTZ-DK basis set (Fe, 9s8p6d3f2g),[32] which can handle the 3s3p semicore correlation of Fe well. For the six atoms of the immediate coordination sphere, we employed a triple-ζ cc-pvTZ-DK basis set (O, N, 4s3p2d1f; S, 5s4p2d1f),[33] while for the rest of the system we used the double-ζ cc-pvDZ-DK basis set (C, 3s2p1d; H, 2s1p).[33] The total number of basis functions is 631.

Cholesky decomposition (CD) techniques[34] were used during the CASPT2/CASSCF calculations with a well-tested threshold of 10^{-4} (see Table S1 in the SI) to reduce the computational time and disk storage requirements. All valence electrons plus the 3s and 3p electrons of iron were correlated in the CASPT2 calculations. To bracket more critically the results, the CASPT2 calculations used two types of zero order Hamiltonians: the standard IPEA zero-order Hamiltonian, using an orbital energy shift correction by the gap between ionization potential (IP) and electron affinity (EA) of active orbitals, i.e., IPEA shift = 0.25. This latter procedure is suggested in MOLCAS, as a method of choice, for reducing systematic errors of calculated gaps between states having different numbers of unpaired electrons.[35] However, to the best of our knowledge, there are still no systematic calibrations to show that this IPEA shift value provides better results for other cases, i.e., in isogyric processes where the number of unpaired electrons is kept constant. Therefore, we used also a second zero-order Hamiltonian (IPEA shift = 0), which is the original formulation, and applied it for some triradicaloid states with three unpaired electrons such as the a_{1u} singly occupied states, quartet Fe^V states, and sulfur-based triradicaloid states. An imaginary level shift of 0.1 au was used in the CASPT2 calculation to avoid the intruder state problem.[36]

As already noticed by Radoń et al. in their gas phase calculation,[14] state-average CASSCF orbitals could be far from optimal for any of the states involved. So in this work, wherever possible, we used state-specific calculations to compute one state at a time. For higher excited states for which state-specific calculations were not possible due to root flipping, we performed state-average calculations followed by multistate (MS) CASPT2 calculations[37] to improve the possible deficiency of the state-average reference CASSCF wave function.

The used QM region and active orbitals are shown in Figure 1. The 14 orbitals depicted in the figure formed an active space of 15 electrons distributed in 14 orbitals, labeled as (15,14). The active space includes for the first time all iron 3d orbitals, and this enables us to explore all the possible configurations within the iron valence 3d shell. The $4d_{xy}$ orbital and $4d_{yz}+3p_y$ and $4d_{xz}+3p_x$ are orbitals accounting for the so-called "double-shell" effect[12] in Cpd I and are found to be important for a balanced treatment of states wherein an electron is excited out of the $3d_{xy}$ (as in the case of the pentaradicaloid state) or Fe−O π type orbitals (as in the case of doublet Fe^V states). The latter two orbitals were first included in the active space by Neese and Thiel et al. in their DDCI2-Q/MM calculation for Cpd I in $P450_{cam}$.[13] These two orbitals are seen in Figure 1 to be combinations of iron 4d and oxygen 3p orbitals rather than pure iron 4d orbitals because of strong covalent interaction between Fe and O of the iron-oxo moiety. All three of the above empty orbitals for the "double-shell" effect were also found to have an additional beneficial effect of stabilizing the active space used here.

Using CASSCF(15,14) calculations (see Table S2−S3 in SI), we found that the states characterized by singly occupied porphyrin a_{1u} or a_{2u} orbitals can be studied separately with active spaces that exclude the corresponding a_{2u} or a_{1u} orbitals. This finding enabled us to use two (13,13) active spaces generated from (15,14) by moving the a_{1u} or a_{2u} orbital out of active space and thereby studying more economically the a_{2u} or a_{1u} singly occupied states. For sulfur-based radicaloid states, it was necessary to introduce two sulfur-based σ_S and π_S orbitals, which are the sulfur lone pair orbitals, the former one pointing toward iron and the latter being perpendicular to the Fe−S−H plane. These two orbitals are not shown in Figure 1 since they are only used to calculate the sulfur-based radicaloid states. The active

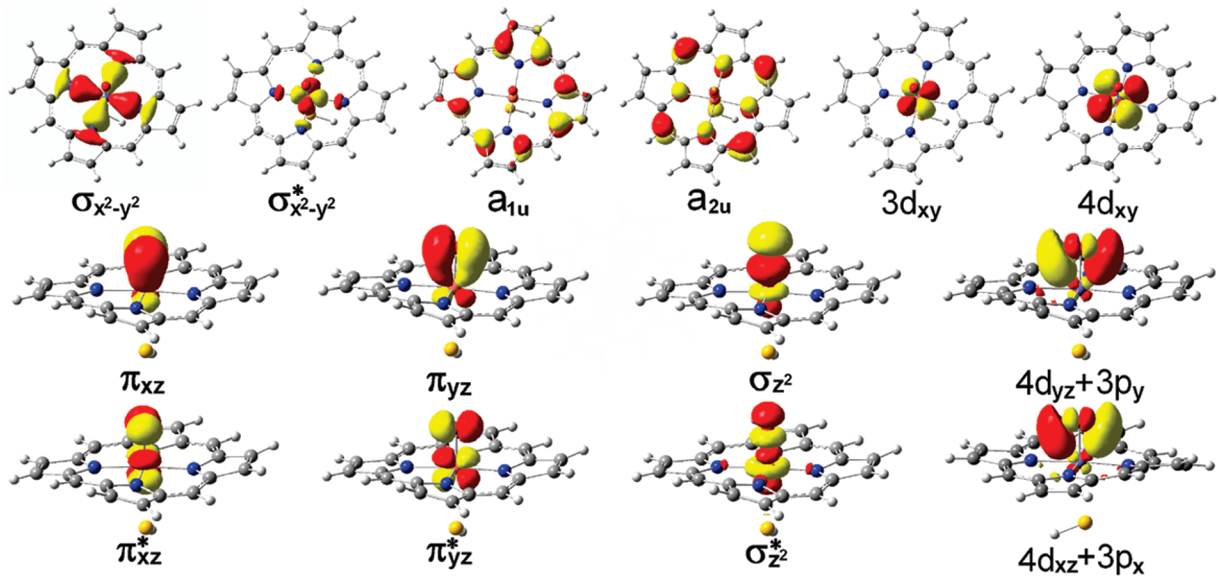


Figure 1. Active orbitals used to generate active space (15,14). The contour value is ±0.05 e/au^3.

space employed for the sulfur-based radicaloid states Π_S/Σ_S was generated from the one shown in Figure 1, by replacing the a_{1u} and a_{2u} orbitals with sulfur-based π_S/σ_S orbitals.

For simplicity, the various states are conventionally labeled by the orbital that accommodates the free radical; we use superscripts to represent spin multiplicity of the state, and in parentheses we add the formal oxidation state of iron; e.g., $^2A_{2u}(Fe^{IV})$ represents the triradicaloid Fe^{IV} state in Scheme 1, while $^4\Delta_{xy}(Fe^{V})$ represents the quartet Fe^{V} state in which iron $3d_{xy}$ (δ) orbital is singly occupied. If there is more than one state for a given label, we add a serial number before the label to indicate energy ordering at the CASPT2/MM level, e.g., 2 $^2A_{2u}(Fe^{IV})$ represents the second Fe^{IV} doublet state, in which porphyrin a_{2u} orbital is mainly singly occupied. The singly occupied orbitals are specified in the fifth columns of Tables 2 and 5. Schemes 2 and 3 further show the simple orbital diagram for these states.

Results and Discussion

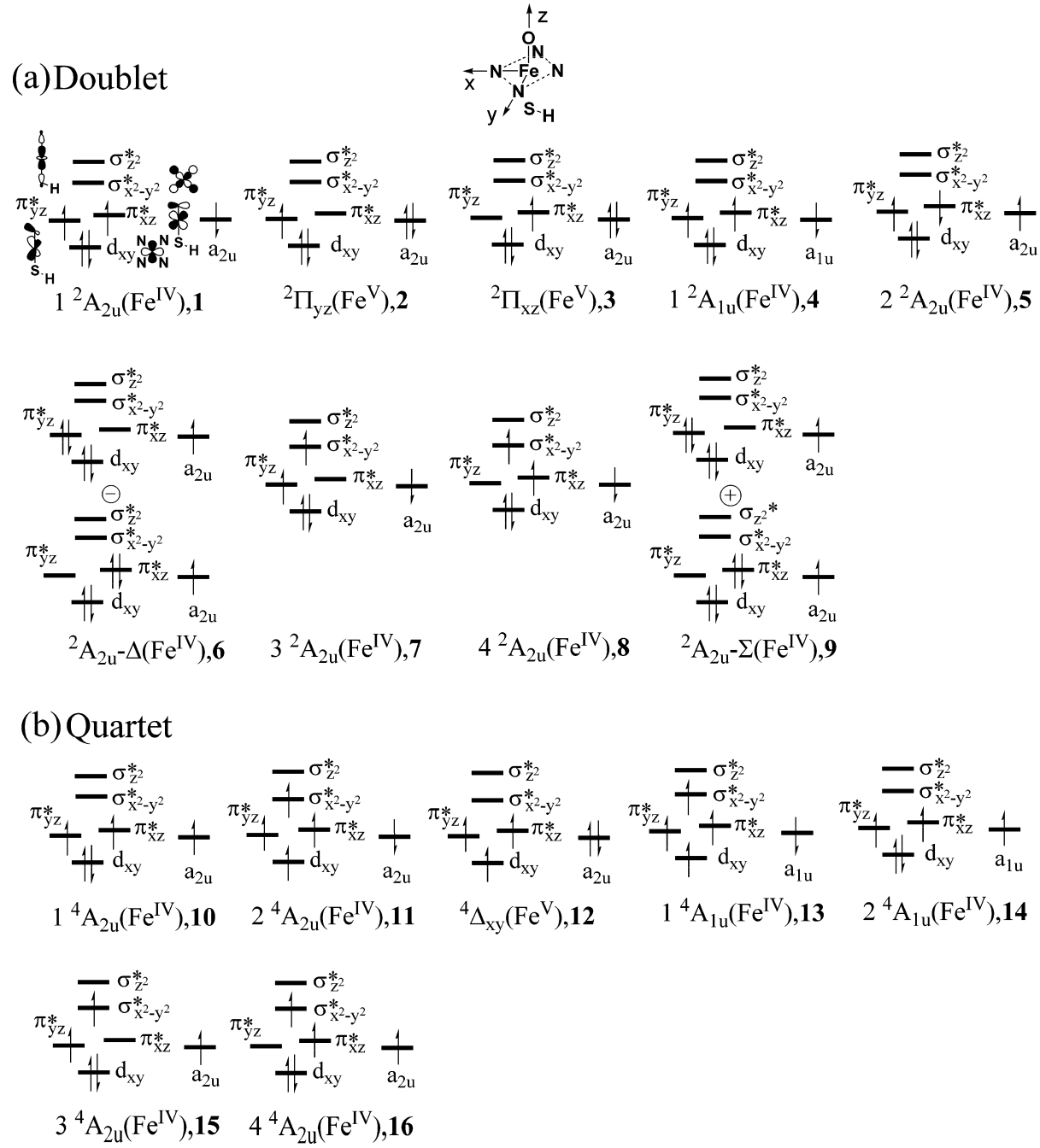
As was shown in the previous theoretical work on model nonheme iron-oxo species,[17a] some geometric parameters such as equatorial Fe−N bond distances are quite important for the energy gap of different states. Therefore, we first summarize the key bond distances of Cpd I obtained from B3LYP/MM calculations for CPO and $P450_{cam}$ in Table 1. It can be seen that, for CPO, where experimental data of extended X-ray absorption fine structure (EXAFS) measurement are available,[38] the bond distances all agree well with the experimental results, except for the Fe−S bond distance, which is about 0.1 Å longer in the calculated geometry. From previous calculations for iron-oxo species as well as our own test calculation at the DFT/MM level (see Table S4 in SI), this geometric difference is not likely to affect the energetic levels of the states we are interested in.

Our main results at the CASPT2/CASSCF/MM level are summarized in Tables 2−5 for Cpd I of CPO and $P450_{cam}$. We calculated 11 doublet and 9 quartet states for each enzyme in Tables 2 and 5. Unless specified, all the relative energy data that we used from ab initio multireference calculations refer to the ones at the CASPT2/MM level. Identification of the various states can be aided by the orbital occupancies in the tables and in Schemes 2 and 3. The full set of results is summarized in the Supporting Information document, while below we present only the key results.

A. Triradicaloid Fe^{IV} Doublet and Quartet States. As seen from Table 2 and Scheme 2, for CPO and $P450_{cam}$, our lowest doublet and quartet triradicaloid states (**1** and **10**) at the CASPT2/MM level correspond, with one exception, to the lowest two states in DFT/MM calculations, in which porphyrin a_{2u} and two Fe−O π^* orbitals are singly occupied, with good accord between the two sets of results. Thus, in agreement with B3LYP/MM, the quartet state, 1 $^4A_{2u}(Fe^{IV})$, is very slightly higher (by 0.6 to 0.9 kcal/mol) in energy than the corresponding doublet state, 1 $^2A_{2u}(Fe^{IV})$, which is stabilized slightly by the antiferromagnetic coupling $S = 1$ iron-oxo ferryl unit and $S = 1/2$ porphyrin radical in the thiolate-ligated Cpd I system. This suggests that DFT and DFT/MM are quite reliable in describing the quartet and doublet ground states of Cpd I. The second triradicaloid doublet state 2 $^2A_{2u}(Fe^{IV})$ (**5**), which involves singlet pairing of the electrons in the two π^* orbitals, and a singly occupied a_{2u}, was located to lie about 18 kcal/mol higher above the ground state **1**, which is comparable with gas phase B3LYP calculations reported before for Cpd I[39] and nonheme iron-oxo species.[40] As such, these states for either CPO or P450cam are not likely to be accessible for reactivity, since most calculated barriers for hydroxylation and epoxidation are lower than this value.[41] We also calculated two other triradicaloid quartet states 3 $^4A_{2u}(Fe^{IV})$ and 4 $^4A_{2u}(Fe^{IV})$ (**15** and **16**) where one of the two electrons in two π^* orbitals is excited to the Fe−N antibonding $\sigma^*_{x^2-y^2}$ orbital. Their high relative energy (more than 26 kcal/mol) indicates they are even less likely than state **5** to affect the reactivity of Cpd I.

B. The Pentaradicaloid Fe^{IV} Quartet State. From Table 2, the pentaradicaloid Fe^{IV} quartet state 2 $^4A_{2u}(Fe^{IV})$ (**11**), having five unpaired electrons, is very close (within 2 kcal/

Scheme 2. Schematic Representation of Orbital Occupancies of the Main Electronic Configurations of Various (a) Doublet and (b) Quartet States of Cpd I Calculated in Table 2[a]

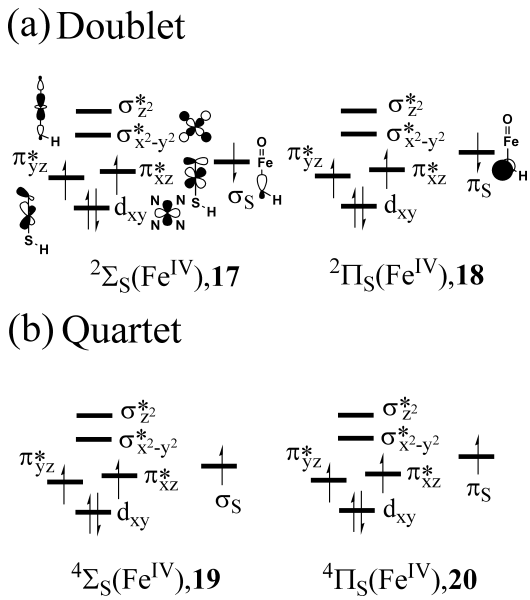


[a] The number in bold after state label is the entry number in Table 2.

mol) to the triradicaloid Fe^{IV} states 1 $^{2,4}A_{2u}(Fe^{IV})$ (**1** and **10**) in energy in both CPO and $P450_{cam}$. We note that there exists also a $^6A_{2u}(Fe^{IV})$ state with a spin-up a_{2u} electron,[8,9,16] which is not computed here. On the basis of the results for the 1 $^{2,4}A_{2u}(Fe^{IV})$ and previous DFT and DFT/MM calculations,[8,9,16] this sextet pentaradicaloid state will lie slightly higher than the quartet state, **11**.

As seen from Scheme 2, the pentaradicaloid Fe^{IV} state **11** differs from the triradicaloid state **1** by excitation of one d_{xy} electron to Fe−N antibonding $\sigma^*_{x^2-y^2}$ orbital. The gap between state **11** and **1** is smaller than the ones provided by the B3LYP/MM calculation of CPO and $P450_{cam}$ where vertical gaps from the ground doublet Fe^{IV} state to this state are both calculated to be about 14−15 kcal/mol at the B3LYP(B2)/MM level. We also tried the recently developed double hybrid functional B2-PLYP,[42] which showed very promising results for improvement of the B3LYP functional in previous extensive calibration calculations.[43] Using a split-valence triple-ζ polarized level basis set (Def2-TZVP[44]), the B2-PLYP/MM calculated vertical gap of 16.9 kcal/mol, between the pentaradicaloid state and triradicaloid state, is close to the B3LYP/MM results in $P450_{cam}$. Thus, CASPT2/MM gives a substantially lower gap than any DFT(functional)/MM method that we tested. In fact, CASPT2/MM

Scheme 3. Schematic Representation of Orbital Occupancies of Main Electronic Configuration of Sulfur Orbital Triradicaloid States of Cpd I Calculated in Table 5[a]



[a] The number in bold after state label is the entry number in Table 5.

Table 1. DFT/MM Calculated and Experimental Key Bond Distances (Å) of the Cpd I Species of CPO and $P450_{cam}$

	B3LYP/MM[a]			exptl[b]		
enzyme	Fe−O	Fe−N_{eq}[c]	Fe−S	Fe−O	Fe−N_{eq}[c]	Fe−S
CPO	1.652	2.026	2.571	1.65	2.01	2.48
$P450_{cam}$	1.652	2.032	2.515			

[a] B1 used for geometry optimization. [b] Ref 38, no experimental data for $P450_{cam}$ are available. [c] Averaged value for four equatorial Fe−N bonds.

predicts that the pentaradicaloid state for P450 is actually the lowest lying state. However, we must note that the original zero-order Hamiltonian of CASPT2 is known to overestimate the stability of states with more unpaired electrons.[35b] The current zero-order Hamiltonian (IPEA shift = 0.25) used here, designed to minimize this deficiency, has been obtained from a fitting of CASPT2 calculations for some small molecule vis-à-vis experimental data,[35a] and one cannot exclude the possibility that it still favors the pentaradicaloid state over the triradicaloid state for Cpd I.[45] Thus, due to the accuracy limit of the CASPT2 method (about 0.2−0.3 eV) and the proximity of the two states, we still cannot determine the accurate gap between the pentaradicaloid Fe^{IV} quartet state **11** and the ground doublet triradicaloid Fe^{IV} state **1**. What we can deduce from the results is that **11** may be closer to **1** than the DFT/MM datum. We note that previous ab initio multireference correlated treatments for other nonheme iron-oxo species also gave a similar energetic proximity of corresponding $S = 1$ and $S = 2$ ferryl units.[8,17]

Table 3. The Mulliken Charges of Fe^V and Fe^{IV} States Calculated at the CASPT2/MM Level for Cpd I of $P450_{cam}$

		Mulliken charge			
state	entry	Fe	O	Por	S
1 $^2A_{2u}(Fe^{IV})$	**1**	1.65	−0.34	−0.56	−0.86
$^2\Pi_{yz}(Fe^V)$	**2**	1.72	−0.13	−0.90	−0.81
$^4\Delta_{xy}(Fe^V)$	**12**	1.78	−0.19	−0.87	−0.84

Table 4. The Mulliken Spin Population of Some Low-Lying Fe^V and Fe^{IV} States Calculated at the CASSCF/MM Level for Cpd I of CPO and $P450_{cam}$

		CPO				$P450_{cam}$			
state	entry	Fe	O	Por	S	Fe	O	Por	S
1 $^2A_{2u}(Fe^{IV})$	**1**	0.76	0.59	−0.33	−0.02	0.77	0.57	−0.33	−0.01
$^2\Pi_{yz}(Fe^V)$	**2**	0.89	0.14	−0.03	0.00	0.78	0.25	−0.03	0.00
$^2\Pi_{xz}(Fe^V)$	**3**	0.80	0.23	−0.03	0.00	0.90	0.13	−0.03	0.00
1 $^4A_{2u}(Fe^{IV})$	**10**	1.17	0.85	0.93	0.05	1.15	0.87	0.92	0.06
2 $^4A_{2u}(Fe^{IV})$	**11**	3.27	0.20	−0.45	−0.02	3.27	0.20	−0.45	−0.02
$^4\Delta_{xy}(Fe^V)$	**12**	2.09	0.99	−0.08	0.00	2.11	0.98	−0.09	0.00

Table 2. CASPT2/MM Relative Energies (kcal/mol), Occupancies of Main Configurations, and Weights (%) of the Main Configurations for Doublet and Quartet States of Cpd I of CPO and $P450_{cam}$ Shown in Scheme 2

		CASPT2/MM[a]		occupancy of main	
state	entry	CPO	$P450_{cam}$	configuration	weight (%)[b]
1 $^2A_{2u}(Fe^{IV})$[c]	**1**	0.0/0.0	0.0/0.0	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^1$	83/83
$^2\Pi_{yz}(Fe^V)$[d]	**2**	5.6	6.0	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^0(\sigma^*_{x^2-y^2})^0(a_{2u})^2$	70/71
$^2\Pi_{xz}(Fe^V)$[d]	**3**	5.8	6.7	$(d_{xy})^2(\pi^*_{yz})^0(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^2$	70/70
1 $^2A_{1u}(Fe^{IV})$[d]	**4**	15.6/17.4	17.7/18.8	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{1u})^1$	82/82
2 $^2A_{2u}(Fe^{IV})$[e]	**5**	18.3	18.4	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^1$	77/76
$^2A_{2u}-\Delta(Fe^{IV})$[e]	**6**	19.6	19.4	$(d_{xy})^2(\pi^*_{yz})^2(\pi^*_{xz})^0(\sigma^*_{x^2-y^2})^0(a_{2u})^1 - (d_{xy})^2(\pi^*_{yz})^0(\pi^*_{xz})^2(\sigma^*_{x^2-y^2})^0(a_{2u})^1$	77/76
3 $^2A_{2u}(Fe^{IV})$[e]	**7**	23.2	23.7	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^0(\sigma^*_{x^2-y^2})^1(a_{2u})^1$	66/66
4 $^2A_{2u}(Fe^{IV})$[e]	**8**	25.4	24.6	$(d_{xy})^2(\pi^*_{yz})^0(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^1(a_{2u})^1$	64/63
$^2A_{2u}-\Sigma(Fe^{IV})$[e]	**9**	28.5	27.2	$(d_{xy})^2(\pi^*_{yz})^2(\pi^*_{xz})^0(\sigma^*_{x^2-y^2})^0(a_{2u})^1 + (d_{xy})^2(\pi^*_{yz})^0(\pi^*_{xz})^2(\sigma^*_{x^2-y^2})^0(a_{2u})^1$	57/61
1 $^4A_{2u}(Fe^{IV})$[d]	**10**	0.9/1.2	0.6/0.0	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^1$	83/83
2 $^4A_{2u}(Fe^{IV})$[d]	**11**	0.0	−1.9	$(d_{xy})^1(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^1(a_{2u})^1$	80/80
$^4\Delta_{xy}(Fe^V)$[d]	**12**	2.2/8.0	2.2/7.4	$(d_{xy})^1(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^2$	79/78
1 $^4A_{1u}(Fe^{IV})$[d]	**13**	15.6	16.9	$(d_{xy})^1(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^1(a_{1u})^1$	80/80
2 $^4A_{1u}(Fe^{IV})$[d]	**14**	17.6/19.8	19.2/20.6	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{1u})^1$	83/83
3 $^4A_{2u}(Fe^{IV})$[e]	**15**	26.2	28.6	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^0(\sigma^*_{x^2-y^2})^1(a_{2u})^1$	65/66
4 $^4A_{2u}(Fe^{IV})$[e]	**16**	26.6	29.2	$(d_{xy})^2(\pi^*_{yz})^0(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^1(a_{2u})^1$	65/66

[a] The values after the slash are from the original zero-order Hamiltonian (IPEA shift = 0), while the other values (before slash or without slash) are from the standard IPEA zero-order Hamiltonian (IPEA shift = 0.25). [b] Weight of the shown main configuration state functions (CSFs) of CASSCF wave function as represented in Scheme 2, data shown in the CPO/$P450_{cam}$ pattern. [c] This state is taken as the zero of the relative energy scale. [d] Single state calculation. [e] State-average calculation.

Table 5. CASPT2/MM Relative Energies (kcal/mol), Occupancy of Main Configurations, and Weight (%) of the Main Configurations for Four Sulfur-Based Triradicaloid Quartet and Doublet States of Cpd I of CPO and $P450_{cam}$ Shown in Scheme 3

state	entry	CASPT2/MM[a] CPO	CASPT2/MM[a] $P450_{cam}$	occupancy of main configuration	weight (%)[b]
$1\ ^2A_{2u}(Fe^{IV})$[c]	**1**	0.0	0.0	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^1(\sigma_S)^2(\pi_S)^2$	83/83
$^2\Sigma_S(Fe^{IV})$	**17**	14.8	26.7	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^2(\sigma_S)^1(\pi_S)^2$	82/82
$^2\Pi_S(Fe^{IV})$	**18**	25.7	29.9	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^2(\sigma_S)^2(\pi_S)^1$	82/82
$^4\Sigma_S(Fe^{IV})$	**19**	18.0	28.4	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^2(\sigma_S)^1(\pi_S)^2$	82/82
$^4\Pi_S(Fe^{IV})$	**20**	26.5	30.6	$(d_{xy})^2(\pi^*_{yz})^1(\pi^*_{xz})^1(\sigma^*_{x^2-y^2})^0(a_{2u})^2(\sigma_S)^2(\pi_S)^1$	82/82

[a] The values are from the original zero-order Hamiltonian (IPEA shift = 0) and state-specific calculation. [b] Weight of the shown main CSFs of the CASSCF wave function as represented in Scheme 3, data shown in the CPO/$P450_{cam}$ pattern. [c] This state is taken as zero point in energy.

A technically interesting point is that excluding the "outer" $4d_{xy}$ orbital from the active space leads to overestimation of the stability of state **11** relative to state **1** by about 3 kcal/mol (0.15 eV) and reverses the state ordering. This is the first time that the so-called "double-shell" effect[12] of this $4d_{xy}$ orbital is assessed on the relative energies of the $S = 1$ and $S = 2$ states of ferryl iron-oxo species. The potentially similar effect in the nonheme iron-oxo system remains to be explored.

C. Fe^V States. For doublet Fe^V states $^2\Pi_{yz}(Fe^V)$ and $^2\Pi_{xz}(Fe^V)$ (**2**, **3**) and quartet Fe^V state $^4\Delta_{xy}(Fe^V)$ (**12**), our CASPT2/MM calculations turn out also to be quite different from the DFT (B3LYP) or TDDFT/MM results. As reported before, DFT methods converged for the Fe^V state only in the gas phase,[7] but not in a protein environment.[9] In the gas phase, the B3LYP calculations predict that the Fe^V states of Cpd I are about 16−26 kcal/mol higher than the ground state.[7] However, gas phase calculated gaps of the Fe^V to the ground state depend on the used functional and the identity of the axial ligand.[46] Our own exploration of a P450 Cpd I model showed that hybrid functionals predict larger gaps than the GGA functionals, and all the gaps are significant, 15−26 kcal/mol (see Table S5 in the SI). Furthermore, TDDFT/MM in-protein calculations predict an even larger energy gap for the Fe^V state of Cpd I, i.e., more than about 37 for quartet and 46 kcal/mol for doublet states.[9] However, our CASPT2/MM result shows that the standard IPEA zero-order Hamiltonian places the quartet/doublet Fe^V states above the doublet Fe^{IV} ground state in $P450_{cam}$ by only 2.2/6.0 kcal/mol. Interestingly, application of the original CASPT2 zero-order Hamiltonian leads to a somewhat higher gap of 7.4 kcal/mol for the quartet Fe^V state, **12**.[47] So despite the sensitivity of the result to the choice of zero-order Hamiltonian of CASPT2, the calculated CASPT2/MM values are all significantly lower than those from DFT and TDDFT/MM approaches. Interestingly, a recent CASPT2 calculation for chloroiron corrole (Cor) complex shows that, within 1.5 eV of the $Cor^{+\bullet}Fe^{III}Cl$ ground state, there is no high valent Fe^{IV} state in which the corrole ring is closed-shell.[48] This may represent the different tendencies of the corrole and porphyrin ligands to assume a cation radical state.[49]

From the weight of the main configurations shown in last column of Table 2, it can be seen that most of the states, Fe^V included, have dominantly single-configuration characters (70−80%). Thus, the large discrepancy between DFT and CASPT2 based results for these Fe^V states may seem surprising. Nevertheless, we note that this effect is rooted in the large orbital relaxation that attends the CASSCF calculation of Cpd I Fe^V states compared with the Fe^{IV} states. This in turn leads to large stabilization of the Fe^V states in state-specific calculations compared with the state-average calculations at the CASSCF/MM level (the Fe^V states are found to be 32.2−34.9/15.9−18.3 kcal/mol above the triradicaloid ground Fe^{IV} state for state-average/state-specific calculations of CPO, see Tables S6, S7, and S10, Supporting Information). Thus, as exemplified in Figure 2a using a pair of bonding/antibonding orbitals, $\sigma_{x^2-y^2}/\sigma^*_{x^2-y^2}$, the Fe^V state has larger 3d iron components in the Fe−N/Fe−O bonding orbitals $\sigma_{x^2-y^2}/\pi/\sigma_{z^2}$ and smaller 3d components in the Fe−N/Fe−O antibonding $\sigma^*_{x^2-y^2}/\pi^*/\sigma^*_{z^2}$ orbitals, compared with those in the Fe^{IV} state. This different d contribution is caused by the different number of electrons in Fe−O moieties of the Fe^V and Fe^{IV} states, which exerts different screening effects. As shown schematically in Figure 2b, the energy level of the 3d orbitals of iron is lowered in the Fe^V state and thereby affects the first coordinate sphere Fe−O and Fe−N bonding and antibonding orbitals by increasing the contribution of the iron 3d component in the bonding orbitals, while decreasing this contribution to the antibonding orbitals.

Due to this orbital relaxation, which is missing in TDDFT/MM, it is likely that the latter method underestimates the stabilities of the Fe^V states.[50] Indeed, because of the quite different orbitals on iron in the presence of different oxidation states, the state-average treatment in CASSCF/MM calculations for the Fe^V and Fe^{IV} states of Cpd I is also inappropriate. Such a calculation followed by the CASPT2/MM treatment leads to Fe^V states that are substantially lower (by ca. 10 kcal/mol) than the ground Fe^{IV} triradicaloid state **1** (see Table S6−S7 in SI), which we deem to be very unreasonable. Therefore, here we calculated Fe^{IV} and Fe^V states separately to get the most optimal CASSCF orbital set for each state, and hence a balanced CASPT2 treatment. *This finding also cautions against the use of CASSCF state-average treatment for different oxidation states of transition metal containing systems*, especially when the number of states for each metal oxidation state is different, as in the present case. Using our approach, the reference weights of the CASSCF wave functions for the Fe^V and Fe^{IV} states (see Table S10−S11 in SI) in our final CASPT2 wave functions are almost the same, indicating that the treatments for these two states here are balanced.

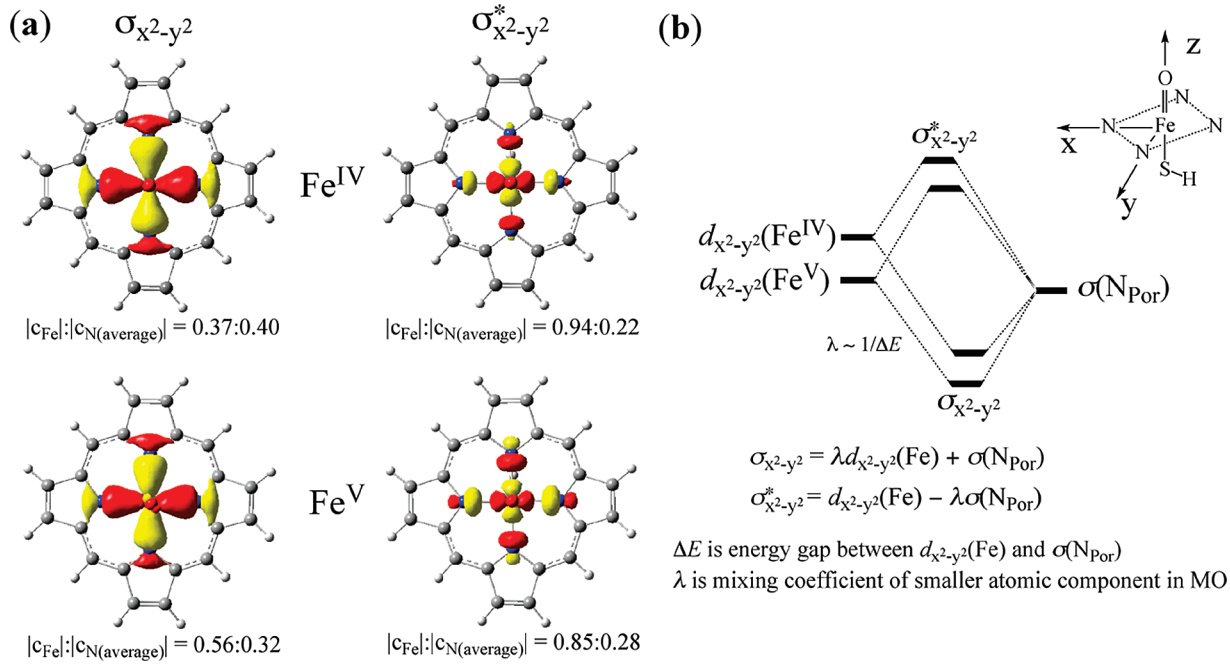


Figure 2. (a) Orbital relaxation of the quartet Fe^V state, $^4\Delta_{xy}(Fe^V)$, compared with the quartet Fe^{IV} state, $1\ ^4A_{2u}(Fe^{IV})$, exemplified by the $\sigma_{x^2-y^2}/\sigma^*_{x^2-y^2}$ bonding/antibonding MOs pair in CPO. The contour value is ±0.05 e/au^3. The data underneath the natural orbitals are the ratios of the absolute values of the Fe and N atomic coefficients in a given MO. (b) A schematic representation of effect exerted by the level shift of an iron d orbital on the orbital relaxation.

The calculated Mulliken charges of the doublet and quintet Fe^V states are compared with those of the Fe^{IV} doublet state at the CASPT2/MM level in Table 3. It can be seen that Fe^V states have smaller negative charges on the O atom of the iron-oxo unit, as would be expected from chemical intuition.

In conclusion, although the CASPT2/MM energy gaps of the Fe^V state are quite low, ≤10 kcal/mol, the variance caused by the two choices of zero-order Hamiltonians does not allow a clear-cut determination of the Fe^V-Fe^{IV} gaps. Perhaps, future calculations with other high level ab initio multireference approaches, such as the SORCI method,[51] which could hopefully be applied to the current large system, will give a more definitive answer to this difficult question of how high the Fe^V states are. Taking the present CASPT2/MM results at face value, the Fe^V states appear to be in principle accessible for affecting the reactivity of Cpd I (see discussion below).

D. Spin Density Distribution in the Fe^{IV} and Fe^V States. The calculated Mulliken spin populations for some of the lowest states in Table 2 are collected in Table 4. The calculated values for two Fe^{IV} triradicaloid states here are close to the previous multiconfigurational calculations.[13,14] The trend among the other states is physically reasonable and can be deduced from the orbital diagrams in Scheme 2. The absolute value of spin density on the thiolate ligand in the triradicaloid states is small ≤0.06, somewhat smaller than in DFT/MM calculations (0.08−0.16).[19a] By comparison, the experimental spin density[3e] for CPO was determined as being smaller than 0.23. No discussion was given[3e] as to how much smaller than 0.23 the actual number is, and it would be interesting to reconsider the experimental data in light of these computed spin densities.

E. $^1\Delta_g$ and $^1\Sigma^+_g$ States Analogous to States of O_2 and a_{1u} Singly Occupied States. As has been noted several times,[52] the FeO moiety of Cpd I has states analogous to those of the O_2 molecule, namely, the $^1\Delta_g$ and $^1\Sigma^+_g$ states. These states of $^2A_{2u}-\Delta(Fe^{IV})$ (**6**) and $^2A_{2u}-\Sigma(Fe^{IV})$ (**9**; in Scheme 2), which involve mixing of two configurations, each having one doubly occupied π^* orbital, are problematic for DFT. We therefore decided to calculate these states as well and find if they are low enough to contribute to the reactivity of Cpd I. As with previous SORCI results for the iron-oxo model complex,[8] our calculated energy gaps (see Table 2) for the doublet states that are analogous to the $^1\Delta_g$ and $^1\Sigma^+_g$ states of the O_2 molecule (**6** and **9**) turn out to be quite large, indicating that they are not likely to become relevant in Cpd I involved reactions.[53]

The importance of the a_{1u} singly occupied states has been debated in the literature quite extensively.[54,55] In our calculation with the standard IPEA zero-order Hamiltonian, the a_{1u} singly occupied states $1\ ^2A_{1u}(Fe^{IV})$, $1\ ^4A_{1u}(Fe^{IV})$, and $2\ ^4A_{1u}(Fe^{IV})$ (**4**, **13**, **14**) were located 15.6−16.7/17.7−18.8 kcal/mol higher than the corresponding a_{2u} singly occupied states $1\ ^2A_{2u}(Fe^{IV})$, $2\ ^4A_{2u}(Fe^{IV})$, and $1\ ^4A_{2u}(Fe^{IV})$ (**1**, **11**, **10**) for CPO/$P450_{cam}$. These substantial gaps compared to the a_{2u} states are larger than a very small value of 1.9 kcal/mol obtained at the DDCI2-Q level.[13] Interestingly, the calculated values are comparable with the TDDFT(B3LYP)/MM value of 12.3 kcal/mol for $P450_{cam}$.[13] In contrast to the Fe^V states above, for the a_{1u} singly occupied state, we do not observe large orbital relaxation phenomenon compared with the a_{2u} singly occupied state; hence the TDDFT results for this state may be more reliable than that for Fe^V states. As observed before,[13,14,56] we notice that the a_{2u} singly occupied states are lower than the corresponding a_{1u} singly occupied states

because of dynamic correlation. Indeed, at the CASSCF/MM level, their energetic levels are reversed compared with the CASPT2/MM situation. As can be seen from Table 2, in contrast to the Fe^V states, the sensitivity of the A_{1u} state energies to the choice of zero-order Hamiltonian of CASPT2 is not significant.

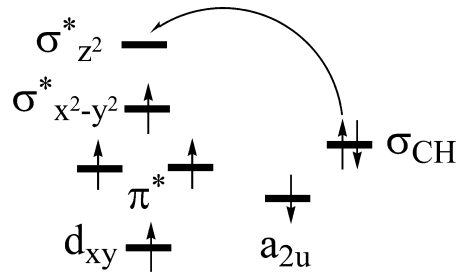
F. Sulfur-Based Radicaloid States. The sulfur-based triradicaloid quartet and doublet states $^{2,4}\Sigma_S(Fe^{IV})$ and $^{2,4}\Pi_S(Fe^{IV})$ shown in Scheme 3, in which the singly occupied porphyrin a_{2u} orbital of the A_{2u} ground states is replaced by singly occupied sulfur σ_S and π_S orbitals, respectively, were found to be very low lying in previous gas phase DFT calculations.[7,57−59] In some cases, depending on the thiolate ligand representation (SCH_3 instead of SH), these sulfur radical states came out as the ground states, and the protein environment was needed to retrieve the A_{2u} ground states.[58] We therefore decided to explore their energy levels using CASPT2/MM calculations. The states are shown in Scheme 3, while the relative energies are collected in Table 5.

Similarly to the case of Fe^V states, here too the state-average orbitals are far from being optimal for the A_{2u} and Σ_S/Π_S states at the same time. Large CASSCF(15,14) calculations (see Table S8 in SI) that include either a_{2u} and one of two sulfur orbitals or two sulfur orbitals in the active space indicate that the A_{2u} state and Σ_S/Π_S state can be calculated separately in active spaces excluding the nearly doubly occupied σ_S/π_S orbital and a_{2u} orbital (with occupancies >1.999), respectively. So we used such a state-specific strategy to calculate one state in each calculation. The results for these states show (see SI for details) that they are at least 14.8 kcal/mol higher above the ground doublet state **1** for CPO at the CASPT2/MM level.[60] This value is higher than the corresponding lowest result (9.6 kcal/mol) of the previous CASPT2 gas phase calculation,[14] clearly indicating that the protein environment causes the sulfur-based triradicaloid state to be less favored than the porphyrin-based triradicaloid state of Cpd I. Another interesting difference between CASPT2 gas phase and CASPT2/MM calculations for the Σ_S/Π_S state is that our CASPT2/MM calculation predicts that the former state is lower in energy than the latter, while in the gas phase the state ordering is opposite. This difference may be caused by the larger stabilization of the π_S orbital compared with the σ_S orbital, due to interactions with the protein environment caused by the different orbital directionalities. The main stabilizing factor for the σ_S orbital comes from the iron ion and it is present in the gas phase already, while for the π_S orbital, the protein may have additional stabilization through an amidic hydrogen-bonding-type interaction (NH···S) between the sulfur atom and the peptide bonds (HN−CO) of the proximal residues around the proximal cysteine ligand. We note that these sulfur-based radicaloid states were predicted to be 60−70 kcal/mol lower than the a_{2u} triradicaloid state at the CASSCF level in the previous gas phase calculation,[14] but here within the protein environment, the sulfur and a_{2u} triradicaloid states of Cpd I are already comparable at the CASSCF/MM level, in accord with previous analysis.[7] This situation in protein makes the multistate strategy less crucial here than that in the gas phase (see Table S9 in SI).[14]

Another noteworthy feature in Table 5 is that the differences between the corresponding relative energies of sulfur-based triradicaloid states of CPO and $P450_{cam}$ are quite large, which is in contrast with the similar state gaps for the other states in Table 2. The gaps in CPO are significantly smaller than the corresponding ones in $P450_{cam}$, especially for the Σ_S state. This result indicates that the relative stability of sulfur-based triradicaloid states can be quite enzyme-dependent. It is generally accepted that CPO has a more polar pocket than the one possessed by $P450_{cam}$,[1,2] which was also confirmed by the previous QM/MM analysis.[19a] But our results show that at least for the region near the sulfur atom, the accumulation of the negative charges on S in state **1** compared with the sulfur-based triradicaloid states (**17**−**20**) is less favored energetically in CPO than that in $P450_{cam}$. This result is quite surprising since it is usually considered that the polar environment should generally favor the electronic state with more charge separation and localization. The most probable explanation for these very different relative energies of the Σ_S states in the two enzymes could be due to the Fe−S bond distance, which is about 0.06 Å shorter in $P450_{cam}$ than that in CPO at the QM/MM level. This shorter bond stabilizes the σ_S orbital of S (but less for π_S) and thereby increases the corresponding state energy gap for $P450_{cam}$.

G. A Brief Discussion on the Potential Impact of the Cpd I States on Reactivity. The above CASPT2/MM results reveal that Cpd I is a remarkable reagent having more than 20 states jammed within 30 kcal/mol. Considering that the computed barriers for H abstraction range between 6−23 kcal/mol,[1g] depending on the substrate, it might be expected that many of the states will contribute to the reactivity of Cpd I by affecting the transition state character, and some may even be involved directly. In this last respect, CASPT2/MM shows that in addition to the consensual triradicaloid states, $1\ ^{2,4}A_{2u}(Fe^{IV})$, there may be four additional low-lying states. These are the pentaradicaloid, $2\ ^4A_{2u}(Fe^{IV})$, and the perferryls, $^2\Pi_{yz/xz}(Fe^V)$ and $^4\Delta_{xy}(Fe^V)$. In the $^6A_{2u}(Fe^{IV})$ state (not calculated in this work), the a_{2u} radical being ferromagnetically coupled with a $S = 2$ Fe−O ferryl unit should also be close in energy to the pentaradicaloid $2\ ^4A_{2u}(Fe^{IV})$ state. We cannot say with certainty that the CASPT2/MM energy gaps are converged with respect to the active space and basis set used, but if we take the present results at their face values, then the energy gaps of the perferryl and pentaradicaloid states relative to the $1\ ^{2,4}A_{2u}(Fe^{IV})$ states are well below most of the barriers calculated so far for the triradicaloid states.[1f,g] One should then consider that the pentaradicaloid and perferryl states should be energetically available along the reaction pathways of Cpd I. In such a putative scenario, if the crossover probabilities from the triradicaloid to the pentaradicaloid and perferryl states were nonzero, these latter states, being so low in energy and perhaps also possessing small barriers, would have led to MSR and brought about a significant enhancement of reactivities of the P450 and CPO Cpd I species. However, with one exception in model systems,[10] all the other LFP generated Cpd I species of P450 and synthetic models are reactive but not overwhelmingly so.[6] *This suggests that there are no additional low-lying*

Scheme 4. Schematic Representation of Increase of the Exchange Interaction during Electron Redistribution in Hydrogen Abstraction of C−H Bond by Quartet Pentaradicaloid State



states that participate in a direct way in the normal reactivity nascent from the 1 $^{2,4}A_{2u}(Fe^{IV})$ states. However, if the present CASPT2/MM results are reliable, then the Fe^{V} and the 2 $^{4}A_{2u}(Fe^{IV})$ states may affect reactivity if they could be accessed directly, as suggested in recent papers for the perferryl states.[6,18] Will these state-specific reactions indeed be so much faster than the ground state's reaction? This is still a question. Let us then discuss some prospective features of these state-specific reactivities.

Reactivity Patterns of the Pentaradicaloid States. As seen from the main configuration of the pentaradicaloid state (Scheme 2), the number of unpaired electrons on iron-centered orbitals changes from 2 to 4 compared with the triradicaloid state **1**. This brings more exchange interaction for the pentaradicaloid state relative to the triradicaloid states[16] and counteracts the orbital gap due to the d_{xy} to $\sigma^*_{x^2-y^2}$ excitation. This effect increases along the oxidation pathway. Thus, as shown in Scheme 4, for example, during the hydroxylation reaction by the pentaradicaloid state of Cpd I, the number of unpaired electrons on iron increases to five, and hence increasing the number of exchange correlation interactions lowers the hydrogen abstraction barrier from this state relative to the triradicaloid state. Although this is counteracted to some extent by the orbital energy gap, associated with the $a_{2u}\rightarrow\sigma^*_{z^2}$ electron shift, the exchange stabilization wins out and leads to a net stabilization; hence, the pentaradicaloid states would be typified by *exchange-enhanced reactivity*.

The DFT[16] and DFT/MM[9] calculation have shown that hydrogen abstraction barrier on the Fe^{IV} pentaradicaloid state is indeed much lower than that on the Fe^{IV} triradicaloid state. Given the fact that CASPT2/MM predicts the energy of the 2 $^{4}A_{2u}(Fe^{IV})$ state to be virtually degenerate with the ground state, even if its stability is overestimated, the impact of this state on the reactivity of Cpd I has to be seriously considered. We note that a similar conclusion has been reached from DFT/MM and DFT calculations, namely, that, even at these levels, the involvement of the pentaradicaloid states in reactivity is a likely event.[9,16] However, with the exception of a recent tentative suggestion by Groves,[10] there are no experimental data or methods that enable evaluation of the potential contribution of the pentaradicaloid channel to reactivity. Reconsideration of the effect of the pentaradicaloid states on the spectroscopy (e.g., Mössbauer) of the triradicaloid ground state may assist the evaluation of the energy gap between the states.

Reactivity of the Perferryl States. Turning now to the Fe^{V}═O states, we note that reactivities of Fe^{V}═O as well as of Fe^{IV}═O reagents have been studied in the nonheme complexes.[61−72] However, the reactivity of the only characterized perferryl reagent[64] seems to be inferior compared with the most reactive ferryl reagents like [BnTPEN-Fe^{IV}(O)$(O_3SCF_3)]^+$ (BnTPEN, *N*-benzyl-*N*,*N*′,*N*′-tris(2-pyridylmethyl)-1,2-diaminoethane) which can even activate an inert C−H bond like in cyclohexane.[72] In contrast, the reactivity of the putative perferryl reagents, of the L_4Fe^{V}(O)(OH) types, is indeed higher than that of ferryl reagents,[61−63] but since these perferryl complexes have not been characterized, one may question whether or not the experimental trend actually reflects intrinsic perferryl reactivity. Very high reactivity has been reported for the PorMn^{V}O reagents.[73] However, as was argued by Groves[73a] and demonstrated computationally by Groves-Car et al.,[74] Eisenstein et al.,[75] and one of us,[76] the high reactivity of these reagents is due to TSR (two-state reactivity) and involves the higher spin states of these reagents.

In summary, before drawing any conclusions on the putative reactivity of the Fe^{V} and pentaradicaloid states of P450, one would have to ascertain (e.g., by CASPT2/MM) that such state specific reactions are indeed faster than those nascent from the 1 $^{2,4}A_{2u}(Fe^{IV})$ ground states.

Conclusions

The ab initio multireference correlated QM/MM treatment for an open-shell transition-metal-containing biochemical system is highly complex and requires considerable insight and technical control and hence has not been widely used compared with the DFT/MM approach in computational bioinorganic chemistry.[77] However, the increasing number of applications of ab initio multireference correlated methods recently in transition metal containing molecules[8,13,14,17,48,78−92] is naturally leading also to applications in the complex field of metalloenzyme systems. In this work, we performed high level ab initio multireference correlated CASPT2/MM calculations to assess the low-lying states of the important reactive species Cpd I in $P450_{cam}$ and CPO, at a significantly advanced level of state completeness, overlaid with the protein environment effects. This is achieved by using a large active space with all the iron 3d orbitals and some 4d involved orbitals to account for the double-shell effect. Similar with a previous CASSCF/MM application to the oxyheme species in myoglobin,[93] the current CASPT2/MM treatment for the Cpd I system is different in many aspects from the gas phase CASPT2 treatment. This difference underscores the influence of the protein environment on the state description and energy levels in these electron deficient systems. In this respect, *the CASPT2/MM study shows that DFT/MM results are reliable for many of the states studied here*, and with the notable exception of the pentaradicaloid and perferryl states, there is by and large a reasonably good accord between DFT/MM and CASPT2/MM.

The CASPT2/MM calculations predict that in Cpd I in addition to the two nearly degenerate Fe^{IV} quartet and doublet triradicaloid states, as revealed by DFT and DFT/MM methods, both Fe^{IV} pentaradicaloid and Fe^{V} states *are*

possibly accessible in energy. This raises questions whether the Fe^{V} states or Fe^{IV} pentaradicaloid states of Cpd I could contribute directly to the reactivity of Cpd I of P450. Previous DFT/MM calculations suggested that the hydrogen abstraction barrier on the Fe^{IV} pentaradicaloid state is lower than that on the Fe^{IV} triradicaloid state.[9] If these states are really involved in reactions of Cpd I, then a multistate reactivity, rather than the two-state reactivity suggested before,[94] would necessarily become a minimal and a better model to understand Cpd I reactivity in P450. A better assessment of this feature can, however, be made only after ascertaining that the present CASPT2/MM state gaps are converged with respect to increasing the active space and basis set, and by estimation of the barriers of reactions catalyzed by P450 such as hydroxylation at the CASPT2/MM level. Studies along these lines are under way in our research group.

Acknowledgment. We thank Prof. B. O. Roos for helpful discussions. S.S. is supported by an ISF grant (53/09). H.C. thanks the Golda Meir fellowship fund. J.S.S. thanks the financial support from the China Scholarship Council (CSC). W.W. is supported by the Natural Science Foundation of China (No 20533020, 20873106).

Supporting Information Available: Computational procedures and a full set of computational results. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Omura, T.; Sato, R. *J. Biol. Chem.* **1962**, *237*, 1375–1376. (b) Sono, M.; Roach, M. P.; Coulter, E. D.; Dawson, J. H. *Chem. Rev.* **1996**, *96*, 2841–2887. (c) Denisov, I. G.; Makris, T. M.; Sligar, S. G.; Schlichting, I. *Chem. Rev.* **2005**, *105*, 2253–2277. (d) Ortiz de Montellano, P. R. *Cytochrome P450: Structure, Mechanism and Biochemistry*, 3rd ed.; Kluwer Academic/Plenum: New York, 2005. (e) Meunier, B.; de Visser, S. P.; Shaik, S. *Chem. Rev.* **2005**, *104*, 3947–3980. (f) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279–2328. (g) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 947–1017.

(2) (a) Morris, D. R.; Hager, L. P. *J. Biol. Chem.* **1966**, *241*, 1763–1768. (b) Dawson, J. H.; Sono, M. *Chem. Rev.* **1987**, *87*, 1255–1276. (c) Sundaramoorthy, M. Chloroperoxidase. In *Handbook of Metalloproteins*; Messerschmidt, A., Huber, R., Poulos, T. L., Wieghardt, K., Eds.; John-Wiley and Sons: New York, 2001; Vol. 1, pp 233−244.

(3) (a) Rutter, R.; Hager, L. P.; Dhonau, H.; Hendrich, M.; Valentine, M.; Debrunner, P. *Biochemistry* **1984**, *23*, 6809–6816. (b) Palcic, M. M.; Rutter, R.; Araiso, T.; Hager, L. P.; Dunford, H. B. *Biochem. Biophys. Res. Commun.* **1980**, *94*, 1123–1127. (c) Egawa, T.; Proshlyakov, D. A.; Miki, H.; Makino, R.; Ogura, T.; Kitagawa, T.; Ishimura, Y. *J. Biol. Inorg. Chem.* **2001**, *6*, 46–54. (d) Hosten, C. M.; Sullivan, A. M.; Palaniappan, V.; Fitzgerald, M. M.; Terner, J. *J. Biol. Chem.* **1994**, *269*, 13966–13978. (e) Kim, S. H.; Perera, R.; Hager, L. P.; Dawson, J. H.; Hoffman, B. M. *J. Am. Chem. Soc.* **2006**, *128*, 5598–5599. (f) Stone, K. L.; Behan, R. K.; Green, M. T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12307–12310.

(4) (a) Davydov, R.; Makris, T. M.; Kofman, V.; Werst, D. E.; Sligar, S. G.; Hoffman, B. M. *J. Am. Chem. Soc.* **2001**, *123*, 1403–1415. (b) Denisov, I. G.; Makris, T. M.; Sligar, S. G. *J. Biol. Chem.* **2001**, *276*, 11648–11652.

(5) (a) Egawa, T.; Shimada, H.; Ishimura, Y. *Biochem. Biophys. Res. Commun.* **1994**, *201*, 1464–1469. (b) Spolitak, T.; Dawson, J. H.; Ballou, D. P. *J. Biol. Chem.* **2005**, *280*, 20300–20309. (c) Kellner, D. G.; Hung, S. C.; Weiss, K. E.; Sligar, S. G. *J. Biol. Chem.* **2002**, *277*, 9641–9644.

(6) (a) Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2008**, *130*, 13310–13320. (b) Sheng, X.; Zhang, H. M.; Im, S.-C.; Horner, J. H.; Waskell, L.; Hollenberg, P. F.; Newcomb, M. *J. Am. Chem. Soc.* **2009**, *131*, 2971–2976. (c) Wang, Q.; Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2009**, *131*, 10629–10636.

(7) Ogliaro, F.; de Visser, S. P.; Groves, J. T. *Angew. Chem., Int. Ed.* **2001**, *40*, 2874–2878.

(8) Schöneboom, J. C.; Neese, F.; Thiel, W. *J. Am. Chem. Soc.* **2005**, *127*, 5840–5853.

(9) Altun, A.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2007**, *129*, 8978–8987.

(10) Bell, S. R.; Groves, J. T. *J. Am. Chem. Soc.* **2009**, *131*, 9640–9641.

(11) (a) Schöneboom, J. C.; Lin, H.; Reuter, N.; Thiel, W.; Cohen, S.; Ogliaro, F.; Shaik, S. *J. Am. Chem. Soc.* **2002**, *124*, 8142–8151. (b) Schöneboom, J. C.; Cohen, S.; Lin, H.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2004**, *126*, 4017–4034. (c) Bathelt, C. M.; Zurek, J.; Mulholland, A. J.; Harvey, J. N. *J. Am. Chem. Soc.* **2005**, *127*, 12900–12908. (d) Guallar, V.; Baik, M.-H.; Lippard, S. J.; Friesner, R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6998–7002. (e) Guallar, V.; Friesner, R. *J. Am. Chem. Soc.* **2004**, *126*, 8501–8508.

(12) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. *Adv. Chem. Phys.* **1996**, *43*, 219–331.

(13) Altun, A.; Kumar, D.; Neese, F.; Thiel, W. *J. Phys. Chem. A* **2008**, *112*, 12904–12910.

(14) Radoń, M.; Broclawik, E. *J. Chem. Theory Comput.* **2007**, *3*, 728–734.

(15) (a) Green, M. T. *J. Am. Chem. Soc.* **1999**, *121*, 7939–7940. (b) Ogliaro, F.; Cohen, S.; de Visser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2000**, *122*, 12892–12893.

(16) Hirao, H.; Kumar, D.; Thiel, W.; Shaik, S. *J. Am. Chem. Soc.* **2005**, *127*, 13007–13018.

(17) (a) Neese, F. *J. Inorg. Biochem.* **2006**, *100*, 716–726. (b) Berry, J. F.; DeBeer George, S.; Neese, F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4361–4374.

(18) (a) Newcomb, M.; Zhang, R.; Chandrasena, R. E.; Halgrimson, J. A.; Horner, J. H.; Makris, T. M.; Sligar, S. G. *J. Am. Chem. Soc.* **2006**, *128*, 4580–4581. (b) Pan, Z. Z.; Zhang, R.; Fung, L. W.-M.; Newcomb, M. *Inorg. Chem.* **2007**, *46*, 1517–1519. (c) Pan, Z. Z.; Wang, Q.; Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2009**, *131*, 2621–2628. (d) Harischandra, D. N.; Zhang, R.; Newcomb, M. *J. Am. Chem. Soc.* **2005**, *127*, 13776–13777. (e) Watanabe, Y. *J. Biol. Inorg. Chem.* **2001**, *6*, 846–856. (f) Nanthakumar, A.; Goff, H. M. *J. Am. Chem. Soc.* **1990**, *112*, 4047–4049.

(19) (a) Cho, K.-B.; Hirao, H.; Chen, H.; Carvajal, M. A.; Cohen, S.; Derat, E.; Thiel, W.; Shaik, S. *J. Phys. Chem. A* **2008**, *112*, 13128–13138. (b) Lai, W. Z.; Chen, H.; Shaik, S. *J. Phys. Chem. B* **2009**, *113*, 7912–7917. (c) Chen, H.; Hirao, H.; Derat, E.; Schlichting, I.; Shaik, S. *J. Phys. Chem. B* **2009**, *112*, 9490–9500.

(20) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *J. Mol. Struct. (THEOCHEM)* **2003**, *632*, 1–28.

(21) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(22) Smith, W.; Forester, T. R. *J. Mol. Graphics* **1996**, *14*, 136–141.

(23) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100. (b) Lee, C.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789. (c) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (d) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

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

(25) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.

(26) (a) Wachters, A. J. H. *J. Chem. Phys.* **1970**, *52*, 1033–1036. (b) Hay, P. J. *J. Chem. Phys.* **1977**, *66*, 4377–4384. (c) Bauschlicher, C. W., Jr.; Langhoff, S. R.; Partridge, H.; Barnes, L. A. *J. Chem. Phys.* **1989**, *91*, 2399–2411. (d) Stewart, R. F. *J. Chem. Phys.* **1970**, *52*, 431–438.

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

(28) (a) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580–10594. (b) de Vries, A. H.; Sherwood, P.; Collins, S. J.; Rigby, A. M.; Rigutto, M.; Kramer, G. J. *J. Phys. Chem. B* **1999**, *103*, 6133–6141. (c) Sherwood, P.; de Vries, A. H.; Collins, S. J.; Greatbanks, S. P.; Burton, N. A.; Vincent, M. A.; Hillier, I. H. *Faraday Discuss.* **1997**, *106*, 79–92.

(29) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177–2186.

(30) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(31) (a) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89–155. (b) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742–3748.

(32) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107/1−064107/15.

(33) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48–53.

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

(35) (a) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149. (b) Andersson, K.; Roos, B. O. *Int. J. Quantum Chem.* **1993**, *45*, 591–607.

(36) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299–306.

(37) Forsberg, N.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **1997**, *274*, 196–204.

(38) Stone, K. L.; Behan, R. K.; Green, M. T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16563–16565.

(39) de Visser, S. P.; Ogliaro, F.; Harris, N.; Shaik, S. *J. Am. Chem. Soc.* **2001**, *123*, 3037–3047.

(40) Hirao, H.; Kumar, D.; Que, L., Jr.; Shaik, S. *J. Am. Chem. Soc.* **2006**, *128*, 8590–8606.

(41) (a) Shaik, S.; Kumar, D.; de Visser, S. P. *J. Am. Chem. Soc.* **2008**, *130*, 10128–10140. (b) de Visser, S. P.; Kumar, D.; Cohen, S.; Shacham, R.; Shaik, S. *J. Am. Chem. Soc.* **2004**, *126*, 8362–8363.

(42) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108/1−034116/16.

(43) (a) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401. (b) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys.* **2007**, *126*, 124115/1–034116/15. (c) Schwabe, T.; Grimme, S. *Acc. Chem. Res.* **2009**, *41*, 569–579.

(44) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(45) Kepenekian, M.; Robert, V.; Le Guennic, B. *J. Chem. Phys.* **2009**, *131*, 114702/1–114702/8.

(46) Dey, A.; Ghosh, A. *J. Am. Chem. Soc.* **2002**, *124*, 3206–3207.

(47) Due to the systematic error of the original CASPT2 zero-order Hamiltonian for states with different numbers of unpaired electrons, we did not use the original CASPT2 zero-order Hamiltonian to calculate relative energies of doublet Fe^V states.

(48) Roos, B. O.; Veryazov, V.; Conradie, J.; Taylor, P. R.; Ghosh, A. *J. Phys. Chem. B* **2008**, *112*, 14099–14102.

(49) de Visser, S. P.; Ogliaro, F.; Shaik, S. *Chem.—Eur. J.* **2001**, *7*, 4954–4960.

(50) (a) Neese, F. *Coord. Chem. Rev.* **2009**, *253*, 526–563. (b) Neese, F. *J. Biol. Inorg. Chem.* **2006**, *11*, 702–711.

(51) (a) Neese, F. *J. Chem. Phys.* **2003**, *119*, 9428–9443. (b) Neese, F.; Petrenko, T.; Ganyushin, D.; Olbrich, G. *Coord. Chem. Rev.* **2007**, *251*, 288–327.

(52) (a) Shaik, S.; Danovich, D.; Fiedler, A.; Schröder, D.; Schwarz, H. *Helv. Chem. Acta* **1995**, *78*, 1393–1407. (b) Shaik, S.; Filatov, M.; Schröder, D.; Schwarz, H. *Chem.—Eur. J.* **1998**, *4*, 193–199. (c) Filatov, M.; Harris, N.; Shaik, S. *J. Chem. Soc. Perkin Trans. 2* **1999**, 399–410.

(53) It should be noted that the two main configurations of state **6** can be transformed to the main configuration of state **5** by orbital mixing within the two FeO π^* orbitals (resulting in π^* orbitals rotation around Fe−O axis). So these two states should be nearly degenerate as long as the two FeO π^* orbitals are isotropic around the *xy* plane, as confirmed by the CASPT2/MM results (see Table 2).

(54) (a) Vangberg, T.; Lie, R.; Ghosh, A. *J. Am. Chem. Soc.* **2002**, *124*, 8122–8130. (b) Hirao, H.; Shaik, S.; Kozlowski, P. M. *J. Phys. Chem. A* **2006**, *110*, 6091–6099. (c) Derat, E.; Cohen, S.; Shaik, S.; Altun, A.; Thiel, W. *J. Am. Chem. Soc.* **2005**, *127*, 13611–13621.

(55) (a) Rutter, R.; Hager, L. P. *Biol. Chem.* **1982**, *257*, 7958–7961. (b) Hosten, C. M.; Sullivan, A. M.; Palaniappan, V.; Fitzgerald, M. M.; Terner, J. *J. Biol. Chem.* **1994**, *269*, 13966–13978. (c) Terner, J.; Palaniappan, V.; Gold; Weiss, R.; Fitzgerald, M. M.; Sullivan, A. M.; Hosten, C. M. *J. Inorg. Biochem.* **2006**, *100*, 480–501.

(56) Ghosh, A.; Persson, B. J.; Taylor, P. R. *J. Biol. Inorg. Chem.* **2003**, *8*, 507–511.

(57) Green, M. T. *J. Am. Chem. Soc.* **1999**, *121*, 7939–7940.

(58) Ogliaro, F.; Cohen, S.; Filatov, M.; Harris, N.; Shaik, S. *Angew. Chem., Int. Ed.* **2000**, *39*, 3851–3855.

(59) It should be noted that because the DFT method usually converges to the ground state, and σ_S and a_{2u} orbitals are mixed in the gas phase, the Σ_S states usually cannot be obtained at this level, but the $^{2,4}\Pi_S$ states could.

(60) For the purpose of comparison with previous gas phase results in ref 14, these values are calculated by using the original CASPT2 zero-order Hamiltonian.

(61) (a) Chen, K.; Que, L. *J. Am. Chem. Soc.* **2001**, *123*, 6327–6337. (b) Chen, K.; Costas, M.; Kim, J.; Tipton, A. K.; Que, L. *J. Am. Chem. Soc.* **2002**, *124*, 3026–3035. (c) Costas, M.; Que, L. *Angew. Chem., Int. Ed.* **2002**, *41*, 2179–2181. (d) Mas-Ballesté, R.; Que, L. *J. Am. Chem. Soc.* **2007**, *129*, 15964–15972. (e) Company, A.; Gómez, L.; Güell, M.; Ribas, X.; Luis, J. M.; Que, L., Jr.; Costas, M. *J. Am. Chem. Soc.* **2007**, *129*, 15766–15767. (f) Company, A.; Feng, Y.; Güell, M.; Ribas, X.; Luis, J. M.; Que, L.; Costas, M. *Chem.—Eur. J.* **2009**, *15*, 3359–3362. (g) Que, L. *Acc. Chem. Res.* **2007**, *40*, 493–500. (h) Costas, M.; Mehn, M. P.; Jensen, M. P.; Que, L. *Chem. Rev* **2004**, *104*, 939–986.

(62) (a) Bassan, A.; Blomberg, M. R. A.; Siegbahn, P. E. M.; Que, L. *J. Am. Chem. Soc.* **2002**, *124*, 11056–11063. (b) Bassan, A.; Blomberg, M. R. A.; Siegbahn, P. E. M.; Que, L. *Angew. Chem., Int. Ed.* **2005**, *44*, 2939–2941. (c) Bassan, A.; Blomberg, M. R. A.; Siegbahn, P. E. M.; Que, L. *Chem.—Eur. J.* **2005**, *11*, 692–705.

(63) Quionero, D.; Morokuma, K.; Musaev, G.; Mas-Ballesté, R.; Que, L. *J. Am. Chem. Soc.* **2005**, *127*, 6548–6549.

(64) de Oliveira, F. T.; Chanda, A.; Banerjee, D.; Shan, X. P.; Mondal, S.; Que, L., Jr.; Bominaar, E. L.; Münck, E.; Collins, T. J. *Science* **2007**, *315*, 835–838.

(65) (a) Nam, W. *Acc. Chem. Res.* **2007**, *40*, 522–531. (b) Yoon, J.; Wilson, S. A.; Jang, Y. K.; Seo, M. S.; Nehru, K.; Hedman, B.; Hodgson, K. O.; Bill, E.; Solomon, E. I.; Nam, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1257–1260. (c) Bukowski, M. R.; Koehntop, K. D.; Stubna, A.; Bominaar, E. L.; Halfen, J. A.; Münck, E.; Nam, W.; Que, L. *Science* **2005**, *310*, 1000–1002. (d) Lim, M. H.; Rohde, J. U.; Stubna, A.; Bukowski, M. R.; Costas, M.; Ho, R. Y. N.; Münck, E.; Nam, W.; Que, L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3665–3670.

(66) (a) Collman, J. P.; Chien, A. S.; Eberspacher, T. A.; Brauman, J. I. *J. Am. Chem. Soc.* **2000**, *122*, 11098–11100. (b) Collman, J. P.; Zeng, L.; Decréau, R. A. *Chem. Commun.* **2003**, 2974–2975.

(67) Mekmouche, Y.; Ménage, S.; Toia-Duboc, C.; Fontecave, M.; Galey, J.-P.; Lebrun, C.; Pécaut, J. *Angew. Chem., Int. Ed.* **2001**, *40*, 949–952.

(68) Klopstra, M.; Roelfes, G.; Hage, R.; Kellogg, R. M.; Feringa, B. L. *Eur. J. Inorg. Chem.* **2004**, 846–856.

(69) Nielsen, A.; Larsen, F. B.; Bond, A. D.; Mckenzie, C. J. *Angew. Chem., Int. Ed.* **2006**, *45*, 1602–1606.

(70) Lee, S. H.; Han, J. H.; Kwak, H.; Lee, S. J.; Lee, E. Y.; Kim, H. J.; Lee, J. H.; Bae, C.; Lee, S. N.; Kim, Y.; Kim, C. *Chem.—Eur. J.* **2007**, *13*, 9393–9398.

(71) Lyakin, O. Y.; Bryliakov, K. P.; Britovsek, G. J. P.; Talsi, E. P. *J. Am. Chem. Soc.* **2009**, *131*, 10798–10799.

(72) Kaizer, J.; Klinker, E. J.; Oh, N. Y.; Rohde, J.-U.; Song, W. J.; Stubna, A.; Kim, J.; Münck, E.; Nam, W.; Que, L. *J. Am. Chem. Soc.* **2004**, *126*, 472–473.

(73) (a) Jin, N.; Groves, J. T. *J. Am. Chem. Soc.* **1999**, *121*, 2923–2924. (b) Zhang, R.; Newcomb, M. *J. Am. Chem. Soc.* **2003**, *125*, 12418–12419. (c) Zhang, R.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2005**, *127*, 6573–6582. (d) Song, W. J.; Seo, M. S.; DeBeer George, S.; Ohta, T.; Song, R.; Kang, M.-J.; Tosha, T.; Kitagawa, T.; Solomon, E. I.; Nam, W. *J. Am. Chem. Soc.* **2007**, *129*, 1268–1277.

(74) de Angelis, F.; Jin, N.; Car, R.; Groves, J. T. *Inorg. Chem.* **2006**, *45*, 4268–4276.

(75) (a) Balcells, D.; Raynaud, C.; Crabtree, R. H.; Eisenstein, O. *Chem. Commun.* **2008**, 744–746. (b) Balcells, D.; Raynaud, C.; Crabtree, R. H.; Eisenstein, O. *Inorg. Chem.* **2008**, *47*, 10090–10099. (c) Balcells, D.; Raynaud, C.; Crabtree, R. H.; Eisenstein, O. *Chem. Commun.* **2009**, 1772–1774.

(76) (a) Khenkin, A. M.; Kumar, D.; Shaik, S.; Neumann, R. *J. Am. Chem. Soc.* **2006**, *128*, 15451–15460. (b) Shaik, S.; Hirao, H.; Kumar, D. *Acc. Chem. Res.* **2007**, *40*, 532–542.

(77) (a) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229. (b) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290. (c) Senn, H. M.; Thiel, W. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.

(78) (a) Cramer, C. J.; Wloch, M.; Piecuch, P.; Puzzarini, C.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 1991–2004. (b) Cramer, C. J.; Kinal, A.; Wloch, M.; Piecuch, P.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 11557–11568. (c) Gherman, B. F.; Heppner, D. E.; Tolman, W. B.; Cramer, C. J. *J. Biol. Inorg. Chem.* **2006**, *11*, 197–205. (d) Cramer, C. J.; Tolman, W. B. *Acc. Chem. Res.* **2007**, *40*, 601–608. (e) Malmqvist, P.-Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109/1–204109/10. (f) Gherman, B. F.; Cramer, C. J. *Coord. Chem. Rev.* **2009**, *253*, 723–753. (g) Huber, S. M.; Ertem, M. Z.; Aquilante, F.; Gagliardi, L.; Tolman, W. B.; Cramer, C. J. *Chem.—Eur. J.* **2009**, *15*, 4886–4895. (h) Huber, S. M.; Shahi, A. R. M.; Aquilante, F.; Cramer, C. J.; Gagliardi, L. *J. Chem. Theory Comput.* **2009**, *5*, 2967–2976.

(79) (a) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094. (b) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2006**, *125*, 124303/1–124303/9. (c) Radoń, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824–11832. (d) Vancoillie, S.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 4011–4019. (e) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2008**, *128*, 034104/1−124303/11. (f) Vancoillie, S.; Rulisek, L.; Neese, F.; Pierloot, K. *J. Phys. Chem. A* **2009**, *113*, 6149–6157. (g) Radoń, M.; Scebro, M.; Broclawik, E. *J. Chem. Theory Comput.* **2009**, *5*, 1237–1244. (h) Radoń, M.; Broclawik, E.; Pierloot, K. *J. Phys. Chem. B* **2010**, *114*, 1518–1528.

(80) (a) Ghosh, A.; Taylor, P. R. *Curr. Opin. Chem. Biol.* **2003**, *7*, 113–124. (b) Ghosh, A. *J. Biol. Inorg. Chem.* **2006**, *11*, 712–724. (c) Ghosh, A.; Gonzalez, E.; Tangen, E.; Roos, B. O. *J. Phys. Chem. A* **2008**, *112*, 12792–12798. (d) Ghosh, A.; Taylor, P. R. *J. Chem. Theory Comput.* **2005**, *1*, 597–600.

(81) (a) Herebian, D.; Wieghardt, K. E.; Neese, F. *J. Am. Chem. Soc.* **2003**, *125*, 10997–11005. (b) Fouqueau, A.; Mer, S.; Casida, M. E.; Lawson Daku, L. M.; Neese, F. *J. Chem. Phys.* **2004**, *120*, 9473–9486. (c) Fouqueau, A.; Casida, M. E.; Lawson Daku, L. M.; Neese, F. *J. Chem. Phys.* **2005**, *122*, 044110/1−044110/13. (d) Ray, K.; Weyhermüller, T.; Neese, F.; Wieghardt, K. E. *Inorg. Chem.* **2005**, *44*, 5345–5360. (e) Petrenko, T.; Ray, K.; Wieghardt, K. E.; Neese, F. *J. Am. Chem. Soc.* **2006**, *128*, 4422–4436. (f) Sundararajan, M.; Ganyushin, D.; Ye, S. F.; Neese, F. *Dalton Trans.* **2009**, 6021–6036.

(82) (a) Ordejón, B.; de Graaf, C.; Sousa, C. *J. Am. Chem. Soc.* **2008**, *130*, 13961–13968. (b) Kepenekian, M.; Robert, V.; Le Guennic, B.; de Graaf, C. *J. Comput. Chem.* **2009**, *30*, 2327–2333.

(83) Suaud, N.; Bonnet, M.-L.; Boilleau, C.; Labèguerie, P.; Guihéry, N. *J. Am. Chem. Soc.* **2009**, *131*, 715–722.

(84) (a) Jensen, K. P.; Roos, B. O.; Ryde, U. *J. Inorg. Biochem.* **2005**, *99*, 45–54. Erratum: *J. Inorg. Biochem.* **2005**, *99*, 978−978. (b) Ribas-Ariño, J.; Novoa, J. J. *Chem. Commun.* **2007**, 3160–3162.

(85) (a) Gagliardi, L.; Roos, B. O. *Chem. Soc. Rev.* **2007**, *36*, 893–903. (b) Roos, B. O.; Lindh, R.; Cho, H.-G.; Andrews, L. *J. Phys. Chem. A* **2007**, *111*, 6420–6424. (c) Wang, X. F.; Andrews, L.; Lindh, R.; Veryazov, V.; Roos, B. O. *J. Phys. Chem. A* **2008**, *112*, 8030–8037.

(86) Rode, M. F.; Werner, H.-J. *Theo. Chem. Acc.* **2005**, *114*, 309–317.

(87) Paulovič, J.; Cimpoesu, F.; Ferbinteanu, M.; Hirao, K. *J. Am. Chem. Soc.* **2004**, *126*, 3321–3331.

(88) Smith, J. M.; Sadique, A. R.; Cundari, T. R.; Rodgers, K. R.; Lukat-Rodgers, G.; Lachicotte, R. J.; Flaschenriem, C. J.; Vela, J.; Holland, P. *J. Am. Chem. Soc.* **2006**, *128*, 756–769.

(89) (a) Sears, J. S.; Sherrill, C. D. *J. Phys. Chem. A* **2008**, *112*, 3466–3477. (b) Sears, J. S.; Sherrill, C. D. *J. Phys. Chem. A* **2008**, *112*, 6741–6752. (c) Takatani, T.; Sears, J. S.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 9231–9236.

(90) (a) Rabilloud, F. *J. Chem. Phys.* **2005**, *122*, 134303/1–134303/6. (b) Polestshuk, P. M.; Sem'yanov, P. I.; Ryabinkin, I. G. *J. Chem. Phys.* **2008**, *129*, 054307/1−054307/13.

(91) Wang, B. W.; Wei, H. Y.; Wang, M. W.; Chen, Z. D. *J. Chem. Phys.* **2005**, *122*, 204310/1–204310/8.

(92) Shearer, J.; Dehestani, A.; Abanda, F. *Inorg. Chem.* **2008**, *47*, 2649–2660.

(93) Chen, H.; Ikeda-Saito, M.; Shaik, S. *J. Am. Chem. Soc.* **2008**, *130*, 14778–14790.

(94) Shaik, S.; de Visser, S. P. In *Cytochrome P450: Structure, Mechanism and Biochemistry*, 3rd ed.; Ortiz de Montellano, P. R., Ed.; Kluwer Academic/Plenum: New York, 2005; pp 45−85.

CT9006234

JCTC Journal of Chemical Theory and Computation

Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining

Lanyuan Lu,[†] Sergei Izvekov,[†] Avisek Das,[‡] Hans C. Andersen,[‡] and Gregory A. Voth*,[†]

Center for Biophysical Modeling and Simulation and Department of Chemistry, University of Utah, Salt Lake City, Utah 84112-0850, and Department of Chemistry, Stanford University, Stanford, California 94305

Received December 2, 2009

Abstract: The multiscale coarse-graining (MS-CG) method obtains CG interactions from atomistic configurations, as demonstrated previously for a variety of soft matter and biological systems. In this article, recent advances in MS-CG algorithms are described, and a recently developed computer program MSCGFM for MS-CG calculations is introduced. The algorithms enhance the efficiency and stability of MS-CG computations, and these algorithms are incorporated into the MSCGFM program. As a result of these efforts, MS-CG calculations on large scale systems such as peptide and proteins can become tractable, and the numerical stability of solutions for ill-posed MS-CG problems can be regularized efficiently. Various parallelization strategies are also discussed.

1. Introduction

Molecular dynamics (MD) simulations with conventional empirical force fields such as Charmm,[1,2] Amber,[3] OPLS,[4] and Gromos[5] are widely used as a computational tool to study soft matter and biophysical systems. A standard all-atom force field typically contains parameters for molecular models with one point mass representing each atom or heavy atom. In typical atomistic MD simulations, observable time scales range from picoseconds to a few microseconds, and tractable system sizes are on the scale of nanometers. However, due to the temporal and spatial limitations of simulations based on all-atom resolution models, many processes in biology are not able to be modeled using available force fields.[6,7] In these cases, coarse-grained (CG) modeling is a powerful tool to study phenomena that involve larger time and length scales.[8] Here, we consider CG models that, like typical atomistic molecular models, regard the system as a set of point masses, called "CG sites". Each CG site corresponds to one or more of the atoms on the same molecule in an atomistic model. It is common to construct such CG models so that most molecules are represented by fewer sites than its number of atoms in the atomistic model. In some cases, some of the molecules, e.g., solvent molecules, have no sites at all. CG models have fewer degrees of freedom than the corresponding atomistic system and hence provide a lower resolution description of the molecular system.

There are numerous ways to develop CG models for different systems. Some approaches tune the parameters of the CG model to reproduce selected physical properties of particular systems. For example, in the MARTINI CG force field,[9] developed by Marrink and coworks and implemented for lipid bilayer[9] and membrane-protein systems,[10] parameters are chosen to reproduce the experimental partitioning free energy values for various components. Another category of methods is based on a strategy in which the low resolution CG model is developed using data from a high resolution model, e.g., an all-atom MD simulation. The advantage of this type of CG approach is that it is able to utilize existent and mature all-atom force fields. In such approaches, an ensemble of atomistic configurations is first generated by performing simulations of the high resolution model, and the low resolution model is then parametrized to reproduce certain properties of the high resolution ensemble. The key issue here is which properties to choose to link the models of different resolutions. In the inverse Monte Carlo[11] or

* Author to whom correspondence should be addressed. E-mail: voth@chem.utah.edu.

[†] University of Utah.

[‡] Stanford University.

10.1021/ct900643r
Published on Web 01/27/2010

iterative inverse Boltzmann[12] method, the radial distribution function (RDF) is usually chosen as the target for fitting in order to obtain the CG model parameters. The inverse Monte Carlo method has been applied to many systems. The reader is referred to the article of Murtola et al. for a recent review.[13] Some theoretical discussions of RDF matching can be found in the literature.[14−18] Higher order correlations, such as three-body correlation functions, are not retained by RDF fitting, and including higher order effects can become computationally intractable in the inverse Monte Carlo method.[17] Moreover, CG models that target the RDF do not in principle reproduce thermodynamic properties of the system.[19]

In the multiscale coarse-graining (MS-CG) method,[20−23] the key property that the CG model is constructed to fit is the many-body equilibrium probability distribution function of the positions of the CG sites in the atomistic system. A mapping operator is used to define the location of each CG site in an atomistic system as a function of the positions of the atoms. (For example, if a CG site corresponds to a set of atoms on a molecule, its location might be defined to be the center of mass or the center of geometry of that set of atoms.) The equilibrium distribution of atom positions in the atomistic system then implies an equilibrium distribution of the CG sites *in the atomistic system.* The MS-CG model that corresponds to the atomistic system is defined to be the model whose equilibrium distribution of CG sites is the same as in the atomistic system.

The potential energy function of the MS-CG model satisfies a variational principle. A numerical method for performing variational calculations to obtain approximate representations of that potential energy function has been developed. The method requires, as input data, a large set of configurations generated by computer simulations of the atomistic system at equilibrium. The method also requires a choice of basis functions to be used in the variational calculation. The choice of a linear basis set reduces the numerical work to standard but large scale problems in numerical matrix computation. The method has been successfully applied to a variety of liquids[21,24,25] and biological systems.[20,26−30]

The CG potential defined by the MS-CG method is in fact the many-body potential of mean force of the CG sites in the atomistic system at equilibrium. When a variational approximation for that potential and its gradients are used to simulate the CG model directly, the resulting forces acting on the sites can be regarded as renormalized forces that take into account the effect of the forces associated with the degrees of freedom that were eliminated in going from the atomistic to the CG system.

While the variational principle in MS-CG is strictly defined from the theoretical point of view, the numerical implementation of it is more challenging. The force-matching problem in MS-CG is a typical inverse problem in which one derives the CG model parameters from the data of atomistic forces, through minimizing a defined variational residual. In practice, the least-squares problem that arises thereafter can be written in the form of a linear matrix equation, by choosing appropriate basis functions to ensure the linearity. However, due to the nature of the molecular systems handled by the MS-CG approach, there are some numerical challenges to be overcome in developing highly efficient algorithms for solving the least-squares problem.

The computational aspects of the variational problem are challenging. The numerical problem can be formulated either as that of obtaining the exact solution of a very large set of inhomogeneous linear equations or as that of obtaining a least-squares solution of a much larger overdetermined set of such equations. Most algorithms for very large systems use the latter formulation.

The first challenge in solving the MS-CG least-squares problem arises from the huge dimension of the matrix for the very large complicated systems to which the MS-CG method is to be applied. For example, for problems of interest, the column dimension of the matrix varies from hundreds to tens of thousands. To make matters more challenging, the row dimension of the matrix is proportional to the product of the number of CG sites in an atomistic configuration and the number of atomistic configurations that are used as input data. In order to obtain good statistics for the CG potential and force functions, a large number of all-atom trajectory configurations are needed, which makes storing the matrix in the physical memory of a typical computer a formidable challenge. The whole matrix cannot be stored in memory, which makes it hard to apply many standard least-squares algorithms. Moreover, the enormous problem size can even make it impossible to store the whole matrix on a hard drive and implement some out-of-core algorithms for the matrix equation. The large size of the matrix equation highlights the need to develop "on the fly" algorithms and apply corresponding least-squares algorithms. The sparse nature of the matrix should also be utilized to enhance computation speed and reduce the storage requirements. For example, the matrix in the MS-CG equations typically contains less than 1% nonzero elements for many large scale systems. Therefore, it is clearly desirable to implement algorithms based on sparse matrix data structure for more efficient memory usage and faster computation.

Another challenge in the numerical implementation of the MS-CG equations arises from the fact that, for typical problems of interest, basis sets typically employed, and data sets typically used, the least-squares problem is numerically ill-posed, with the matrix being nearly singular. Especially for complex biomolecular systems, the ill-posed nature of the problem usually creates a degeneracy of solutions that can be affected by statistical noise in the input all-atom configuration data. One way to enhance the stability of the solution is to apply more robust least-squares solvers, such as those designed for ill-posed problems. Moreover, the so-called regularization method can be implemented, which can prevent the overfitting phenomenon and produce solutions with improved smoothness.

In this article, recent advances in the numerical solutions of the MS-CG equations are presented. These include algorithms for reading the all-atom MD trajectories and generating the matrix equation, as well as strategies for solving the least-squares problem. Data structure is closely related to the algorithms, and therefore it will also be discussed. Before these new algorithms were employed, our

original MS-CG program could perform calculations for CG models with only around 10 to 15 types of CG sites due to the limitations of storage space and computation speed. By contrast, the new algorithms in this paper extend the application scope of the MS-CG approach to complex biological systems such as proteins and membranes. All new algorithmic features described here are implemented in the MSCGFM program that has been recently developed for MS-CG calculations.

The structure of this paper is as follows: Section 2 briefly introduces the MS-CG methodology and discusses basis functions in the MS-CG least-squares problem. The choices for dense/sparse matrix data structures are then given in section 3, accompanied by a description of the neighbor list algorithm for generating the matrix equation. In section 4, different algorithms for solving the least-squares problem are discussed, and miscellaneous algorithms are described in section 5. Finally, the last two sections, 6 and 7, contain benchmark calculations and conclusions, respectively.

2. Multiscale Coarse-Graining Methodology

2.1. Summary of the MS-CG Calculations. The most recent state-of-the-art discussion of the MS-CG approach along with its theoretical background can be found elsewhere.[22,23] Here, we will discuss only enough to provide the background for the numerical considerations.

The goal of the MS-CG method is to develop a coarse-grained model for an atomistic system. The atomistic system is in general a molecular system whose configuration can be represented by $\mathbf{r}^n$, which is the collection of position vectors of the n atoms in the sytem. The total number of CG sites on all the molecules is N. The position of site I, denoted $\mathbf{R}_I$, is defined as $\mathbf{R}_I = \mathbf{M}_{\mathbf{R}I}(\mathbf{r}^n)$, where $\mathbf{M}_{\mathbf{R}I}(\mathbf{r}^n)$ is a mapping operator, which is a part of the definition of the CG model.

The CG potential $U(\mathrm{R}^N;\phi)$ is a linear combination of N_D basis functions of an appropriate type, where ϕ is a one-dimensional column vector of D parameters, ϕ_1, ..., ϕ_{N_D}, which are the coefficients of the basis functions. In the CG system, this potential determines the force on each site I

$$\mathbf{F}_I(\mathbf{R}^N;\phi) = -\frac{\partial}{\partial \mathbf{R}_I} U(\mathbf{R}^N;\phi) \tag{1}$$

Both $\mathbf{F}_I(\mathbf{R}^N;\phi)$ and $U(\mathbf{R}^N;\phi)$ are linear functions of the coefficients in ϕ. Obtaining a representation of those functions that can be used in a computer simulation of the CG system is the goal of the MS-CG calculation.

The MS-CG method provides the following prescription for determining the ϕ array and thereby obtaining the CG potential:

1. Perform molecular dynamics simulations of the atomistic system to obtain a large set of n_t configurations $\mathbf{r}^n$ that are representative of a canonical distribution of configurations for a specific temperature and volume. For each configuration, also obtain the force acting on each of the n atoms.

2. Calculate the locations of all the sites for each configuration.

3. Calculate a matrix $\mathbf{F}$ that has $3n_tN$ rows and N_D columns, each of whose elements is calculated from the positions of the N sites in one configuration and from one of the basis functions.

4. Calculate a column vector $\mathbf{f}$ that has $3n_tN$ elements, each of which is calculated from the forces acting on the atoms associated with one site in one atomistic configuration.

5. Determine

$$\arg\min_{\phi}(\mathbf{f} - \mathbf{F}\phi)^T(\mathbf{f} - \mathbf{F}\phi) \tag{2a}$$

which is the column vector ϕ that minimizes the function $(\mathbf{f} - \mathbf{F}\phi)^T(\mathbf{f} - \mathbf{F}\phi)$, which is a quadratic function of the entries of ϕ. Here, T denotes the transpose. This is equivalent to finding the ϕ array that approximately solves the overdetermined set of linear equations

$$\mathbf{F}\phi \cong \mathbf{f} \tag{2b}$$

in a least-squares sense. (The symbol $\cong$ is a reminder that the least-squares solution of the overdetermined set of equations is approximate, rather than exact.)

See refs 22 and 23 for further details.

2.2. Basis Functions. The choice of basis functions for the potential U and force functions $\mathbf{F}_I$ is made on the basis of physical intuition about the contributions that are expected to be important and to be easily representable. Many of them are expressed as functions of simple collective variables, such as the scalar distance between two sites. For example, if x is such a collective variable that is a function of $\mathbf{R}^N$, then a contribution to the potential $U(\mathbf{R}^N;\phi)$ might be of the form

$$\sum_d \phi_d f_d(x) \tag{3}$$

The corresponding contribution to the force $\mathbf{F}_I(\mathbf{R}^N)$ would then contain derivatives of the form $\mathrm{d}f_d(x)/\mathrm{d}x$. Depending on the problem, either the basis functions $f_d(x)$ or $f'_d(x)$, would be represented in some simple form.

In early MS-CG developments of Izvekov and Voth,[21] cubic spline basis functions were implemented for the force functions. Several other types of basis functions including linear spline and delta functions were also tested and compared in the work of Noid et al.[23] Generally speaking, these low-order spline functions are adequate for representing functions of one collective variable. Another set of basis functions was introduced by Das and Andersen to ensure that the force and its spatial derivatives are continuous.[31]

Low-order spline functions have the flexibility required for variational calculations of complicated CG interactions. However, typically, a large number of basis functions are needed for each type of interaction. Consequently, the number of unknowns in eq 2a can be very large, which adds to the computational challenges described in the first section. A natural solution for the problem is to incorporate higher-order polynomial basis functions and reduce the total number of basis functions needed. This has been achieved by applying B-spline functions as basis sets, introduced in the work of Lu and Voth.[29] It has been shown that both the computation time and required memory are reduced by using B-splines in MS-CG calculations, while at the same time

```
DO For all cells
  DO For all CG sites in the cell
    DO For other sites within the same cell
      IF Distance less than cutoff
        Compute forces and fill in matrix elements
      END IF
    ENDDO
    DO For sites in neighboring cells
      IF Distance less than cutoff
        Compute forces and fill in matrix elements
      END IF
    ENDDO
  ENDDO
ENDDO
```

Figure 1. Pseudocode for the neighbor search algorithm with the cell index method.

good accuracy can be retained.[29] Another advantage of B-spline basis functions is that the smoothness of the force function and some of its derivatives can be guaranteed, a desirable feature for filtering statistical noise in force data. However, in practice even with B-spline basis functions, the total number of unknowns in eq 2a can be very large, due to the complexity of realistic systems. Both linear spline and B-spline basis functions have therefore been implemented in the MSCGFM program to deal with different requirements for accuracy and efficiency.

3. Generating the Matrix Equation in Multiscale Coarse-Graining

3.1. Neighbor Search Algorithm. In order to perform the calculations, the vector **f** and the matrix **F** in eq 2a must be constructed. The vector **f** is easily calculated from the atomistic force data that is generated during the original atomistic simulation. The calculation of **F** requires identification of all sets of sites that are close enough to each other that one or more terms in $U(\mathbf{R}^N;\phi)$ can generate forces that act between them. Here, we will discuss only the algorithm for nonbonded interactions between pairs of sites that are close enough to interact.

When the system size is large, the direct search for each nonbonded pair of sites that are within a specific cutoff becomes very inefficient. For such a simple search algorithm, the number of interactions to be computed is $N(N-1)/2$, where N is the total number of CG sites in the system. Fortunately, the cell index method[32,33] widely used in MD simulations can also be implemented here. In this method, the simulation box is divided into a number of cells of a three-dimensional lattice. Each cell has a dimension that is equal to the cutoff distance. When the cell structure is applied, only $13.5NN_c$ interactions are calculated, where N_c is the average number of sites per cell. This is because the neighbor search is only performed for sites within the same cell and neighboring cells. In practice, the cell information is stored in a linked-list data structure, which enables a very efficient neighbor search with the cell index method. Once all the pairs are found, the matrix elements of **F** are calculated. The whole procedure is illustrated in Figure 1 as pseudocode. Although the approach in Figure 1 is specific for pairwise interactions, three-body CG interactions can be treated with a similar procedure. (Interactions that involve more than three sites are usually not practical in MD simulations.) In practice, the MS-CG calculation is dramatically slower without implementation of the cell index method. When the cell index method is used, the computation time for the generation of the **F** matrix step is much less than the time needed for performing the subsequent matrix calculations. Each atomistic configuration that is analyzed adds $3N$ rows to the matrix **F**. Depending on the method used to performed the matrix calculations (discussed later), the partial result for **F** from one configuration or a small block of configurations is converted to an intermediate matrix or a block solution for ϕ. The partial result for **F** is overwritten thereafter.

3.2. Dense and Sparse Matrix Formats. To perform the calculation associated with eq 2a, memory is required for both **F** and **f**, and some additional work space is needed. The memory requirement for storing **F** is by far the largest of the three. As will be shown in the next section, for most MD systems, **F** is a sparse matrix, allowing compact storage via sparse matrix formats. In the MSCGFM code, both dense (standard array data structure) and sparse formats are supported for different least-squares solvers. Although using the sparse matrix format usually saves space, the implementation of a dense matrix data structure makes it straightforward to use a large number of available packages for linear algebra calculations, such as LAPACK.[34] In addition, some equation solving algorithms require operations in dense matrix form, as will be discussed later in this article.

If **F** is very large and sparse, it is more appropriate to apply a sparse matrix format for storing and solving the minimization problem. This is the case for most complex multicomponent systems, especially biomolecular systems in which the number of interactions is very large. For some systems like proteins, the storage of the matrix in dense matrix form becomes impractical, which will be further demonstrated in the next section. It is natural to adopt the conventional Compressed Sparse Row (CSR) format[35] or similar formats since these formats are supported by most sparse linear algebra algorithms and software. In the CSR format, three arrays are needed for the number of nonzero elements in each row, the column indexes of the nonzero elements, and the nonzero element values, respectively. For both the column index array and element value array, both the length and the sequence of the array are fixed, and the memory occupation is continuous, as demonstrated in Figure 2a. These fixed properties of the arrays enable efficient matrix calculations since in these computations the memory access is continuous and the memory range is predefined. However, the CSR format is intrinsically not compatible with the neighbor search approach in Figure 1. In the neighbor searching method, the filling of the matrix elements is random and the total number of nonzero elements is unknown until the whole search procedure is finished. Because the array length is not predefined, allocation of memory then presents a challenge. Although in many programming languages such as C there are functions to dynamically allocate memory, this kind of dynamical allocation usually dramatically lowers computational efficiency in matrix computations. Besides the array length issue, the random filling of the matrix means that every element sequentially

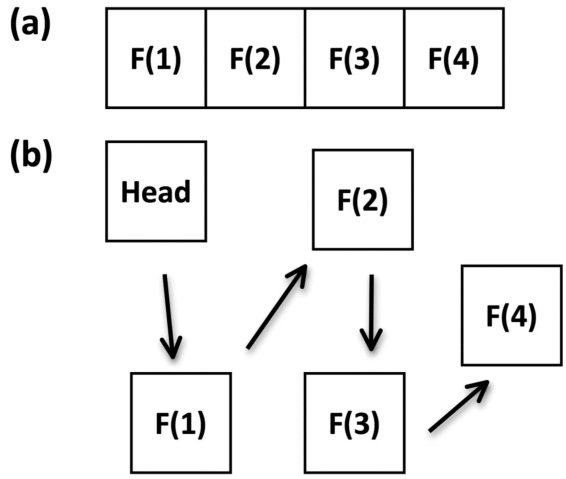


Figure 2. Memory allocations for (a) CSR format and (b) linked list format. In this example, there are four nonzero elements in the sparse matrix.

after the newly filled one must be moved in order to ensure the correct position of each element. This element moving process happens frequently during the neighbor search procedure and can make the whole neighbor search algorithm impractical. Therefore, another sparse matrix format, namely, the linked list format,[36] is implemented for the matrix **F** during the neighbor search. The discontinuous memory occupation of the linked list format is demonstrated in Figure 2b. As shown in that figure, there are no arrays needed in the linked list format, and the matrix elements are stored as scattered "nodes" in memory. For each node, both the element information and the address for the next node are stored. The address information ensures that each node in the linked list can be found sequentially. In addition, there is a "head" node that serves as the starting point of the linked list and stores information like the total number of nonzero elements. Two advantages exist for this linked list data structure in the neighbor search. First, the length of the linked list does not need to be predefined, and it is determined by the address information of the last node. Memory space for each node is created when the new node is generated in the linked list format, which allows efficient usage of computer memory. Second, random insertion of matrix elements becomes computationally efficient since only the address information of the previous node needs to be modified during the insertion. Once the matrix **F** is created after the neighbor search and stored in the linked list format, it can be converted to the conventional CSR format and then is ready to be processed by many sparse linear algebra algorithms. In this equation generation approach with the linked list format, two copies of matrix **F** are created in both linked list and CSR formats. The gain in computation efficiency far outweighs the memory cost of maintaining two copies.

4. Solving the Least-Squares Problem in Multiscale Coarse-Graining

There are a number of established ways to perform the minimization in eq 2a.[37] Due to some special numerical properties of the **F** matrix, several different strategies are discussed here.

4.1. Properties of the Least-Squares Problem. *Matrix Dimension.* Table 1 shows row/column dimensions of the matrix **F** for several simple liquid and biomolecular systems. It is seen that, for simple liquids like water, **F** is rather small, which indicates that eq 2a can be solved rapidly. However, the size of the matrix increases dramatically with system complexity. All biomolecular systems in Table 1 require memory on the order of gigabytes to store the matrix **F**. Note that for the numbers in Table 1 only 10 configurations of the all-atom trajectory are counted. In practice, at least hundreds to tens of thousands of configurations are needed to ensure good statistical force sampling, depending on the complexity of the system. Since the size of **F** is proportional to the number of configurations, in real MS-CG calculations the memory requirement for **F** is usually beyond the hardware limit, even considering sparsity.

Sparsity. Table 1 also shows the percentage of nonzero elements in the matrix **F** for different systems. Traditionally, a matrix is considered to be sparse when the fraction of nonzero elements is less than 5%. For simple liquid systems like water, the matrix is typically not sparse. By contrast, matrices for complicated systems are usually very sparse. This phenomenon is understandable since the number of nonzero interactions is limited for each CG site due to the short range of most of the interactions. For complex systems, the total number of force functions, which corresponds to the column dimension of **F**, is a very large number. Each basis function is typically nonzero only for a small range of its arguments. Therefore, many elements of the **F** matrix are zero. This sparse nature of **F** points to the need to implement sparse linear algebra algorithms, to be discussed later.

Ill-Posed Problems. For many inverse problems, the related least-squares problems are ill-posed. Typically, there are two categories of ill-posed problems, called rank-deficient and discrete ill-posed problems, which are distinguished from one another on the basis of singular value analysis.[38]

In a singular value decomposition (SVD),[39] the matrix **F** can be expressed as product of three matrices:

$$\mathbf{F} = \mathbf{USV}^T \tag{4}$$

where **U** and **V** are orthogonal matrices and **S** is a diagonal matrix with diagonal elements that are called the singular values of the matrix **F**. Conventionally, the singular values S_i are ordered in a nonincreasing fashion, and the condition number is defined as S_{max}/S_{min}. For a rank-deficient problem, there exists a cluster of small singular values, and the gap between small and large singular values is obvious. In this case, the small singular values are usually set to zero in order to regularize the least-squares solution, which is the so-called truncated SVD method.[40]

In the discrete ill-posed problem case, the singular values decrease gradually without any obvious gap. Consequently, various regularization methods need to be implemented in order to balance the residual norm and solution size. Note that common regularization methods used for discrete ill-posed problems can also be applied to rank-deficient cases.

The singular values of a water system and a lipid bilayer system are plotted in Figure 3. From Figure 3, it is seen that the MS-CG problem for the water system is rank-deficient

Table 1. Matrix Dimension and Non-Zero Element Ratio for the MS-CG Calculations for Selected Molecular Systems[a]

system	CG type	n_{row}	n_{col}	$n_{row} \times n_{col}$	memory	nonzero element ratio (%)
1000 water	1	30 000	161	4 830 000	4.6 MB	23
128 lipid (DOPC/DPPC 1:1) + 4000 water	15	173 760	35 031	6 086 986 560	5.7 GB	0.24
A polypeptide + 6256 water	19	199 530	35 058	6 995 122 740	6.5 GB	0.17
T4 Lysozyme + 9739 water	22	297 150	49 292	14 647 117 800	13.6 GB	0.13

[a] For each system, 10 frames of all-atom trajectory are included in the calculation.

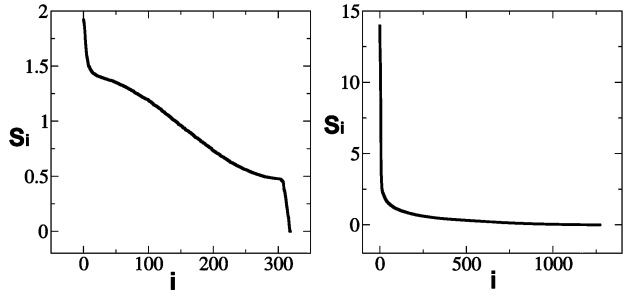


Figure 3. Singular value distributions for (left) a 1000 water system and (right) a 64 DPPC + 3846 water system.

since the smallest singular value 2.9×10^{-16} is by far smaller than any of the other singular values. It is also seen from Figure 3 that the one for the lipid bilayer is a discrete ill-posed problem.

The singular value analysis results can guide one to choose an appropriate least-squares solver and regularization method. Noid et al.[23] point out that, by removing basis functions that are related to unsampled interaction distances, the condition number of **F** can be reduced. Appropriate column rescaling may also reduce the condition number.[23] However, due to statistical noise and redundant information from all-atom configurations,[41] the MS-CG least-squares problem is still typically ill-posed after applying such treatments. In practice, for systems like simple liquids, the MS-CG least-squares problem is usually with full-rank or rank-deficient. For complex systems, rank-deficient and discrete ill-posed problems are often encountered.

4.2. Least Squares Algorithms. From the previous discussion about the matrix size of **F**, it is clear that it is impractical to store the whole **F** in memory, even with sparse matrix formats. Because a large number of all-atom configurations are usually needed, even various out-of-core least-squares algorithms are not able to handle the resultant huge matrix due to hard disk limitations. As a result, two categories of least-squares algorithms have been developed for solving the least-squares problem. The first type of method is based on the "block average" approximation, which is the one used by Izvekov and Voth in the early MS-CG applications.[21] In the second type, the huge size matrix eq 2a is first converted to a smaller equation that is numerically equivalent to eq 2a. Then, the smaller equation is solved by various standard methods. For this category, there are two major algorithms called the normal equation method[39] and the sequential accumulation method,[37] which will be discussed later.

Block Average Algorithm. The basic idea of the block average algorithm is to divide the all-atom trajectory into small blocks of configurations, minimize eq 2a for each block, and obtain the ϕ vector for each block. The last step of the block average algorithm is to calculate the average of the ϕ vectors for the individual block as the final approximate MS-CG solution. Noid et al.[23] pointed out that, if there are significant structural transitions that take place on a time scale that is longer than the block size, there might be systematic error from the block average approach. More theoretical investigation is therefore needed on the applicability of the block average method. However, numerical results from the MS-CG applications to date show that the method is a reasonable approximation for many systems and a discussion can be found in the literature.[23] The block average method is also the only method in which it is possible to use sparse solvers since in this method small problems in the form of eq 2a are directly solved and the sparsity of **F** can be utilized by least-squares solvers. The other algorithms involve matrix transformations that may destroy the sparsity of the relevant matrices.

Normal Equation Algorithm. It is straightforward to show that any solution ϕ of the minimization problem in eq 2a is also a solution of the following equation:

$$\mathbf{F}^T\mathbf{F}\phi = \mathbf{F}^T\mathbf{f} \tag{5}$$

Moreover, the solution of eq 5 is unique if and only if the solution of eq 2a is unique. $\mathbf{F}^T\mathbf{F}$ is an $N_D \times N_D$ matrix (much smaller than **F**, which is $3nN \times N_D$), and so algorithms based on solving eq 5 require less memory than those that deal with eq 2a. The matrix $\mathbf{F}^T\mathbf{F}$ can be constructed directly from the atomistic configuration data without first constructing **F**.

Equation 5 is a set of linear equations that in principle can be solved exactly. Direct methods of solution as well as least-squares methods can be applied. Since the least-squares solver needs to be applied only once to solve eq 5, for small to medium sized systems (i.e., those with up to around 10 types of CG sites), the normal equation algorithm is faster than the block average method. Unlike the case of the block average algorithm, the normal equation algorithm in principle gives the exact solution of the variational problem in eq 2a. However, in matrix computations, the machine error and statistical noise from data need to be considered. The condition number of $\mathbf{F}^T\mathbf{F}$ is the square of the condition number of **F**. Therefore, it is usually not recommended to work with eq 5 for the high condition number problems[37,38] that are often seen in MS-CG applications. Another disadvantage of the normal equation algorithm is that the information of the residual of eq 2a is lost during the normal equation transformation. Although this information is not necessary for a CG force field, it is an important quantity when a regularization method needs to be implemented. The details of regularization will be discussed later in this article.

Sequential Accumulation Algorithm. In the sequential accumulation algorithm introduced by Lawson and Hanson,[37] the precision and condition number issues in the normal

equation algorithm do not exist. The key idea of this algorithm is to convert eq 2a to a smaller least-squares problem as[37]

$$\begin{bmatrix}\mathbf{R}\\0\end{bmatrix}\phi \cong \begin{bmatrix}\mathbf{d}\\e\end{bmatrix} \tag{6}$$

where $\mathbf{R}$ is a upper triangular matrix with dimension $N_D \times N_D$, $\mathbf{d}$ is a vector of N_D length, and e is a real number. This transformation is achieved through a number of QR decompositions in which $\mathbf{F}$ needs to be divided into blocks. Besides obtaining the same set of solutions, eq 6 has some desirable numerical properties. First, it has the same residual as that in the original least-squares problem eq 2a, and the residual is simply the value of e in eq 6. Second, the matrix $\mathbf{R}$ has the same set of singular values as $\mathbf{F}$, and there is no precision loss as in the normal equation algorithm. Due to the above properties, the sequential accumulation algorithm is usually the first choice for very ill-posed problems since the condition number will not increase during the transformation from eq 2a to 6, and various regularization methods can be easily implemented. The memory requirement and computational efficiency of this algorithm highly depend on the block size during the transformation. Generally speaking, the smaller the block size, the smaller the memory space required and the more computation time needed. During the transformation, memory is required for both $\mathbf{R}$ and the block of $\mathbf{F}$. Typically, the sequential accumulation algorithm is slower than the normal equation algorithm.

4.3. Least Squares Solvers. There are a number of standard solvers for solving the least-squares problems, and the choice of solvers largely depends on the availability of optimized software. Many dense solvers such as Gaussian elimination are not robust for problems whose matrix does not have full rank. Thus, they cannot be chosen to solve the MS-CG problem. Two common dense solvers for rank-deficient problems are SVD and QR decomposition with pivoting.[39] For dense matrix algorithms like the normal equation algorithm and sequential accumulation algorithm, the solver is used only once during the whole MS-CG calculation, and the speed of the solver is not very important. Therefore, the SVD method is implemented in this case, although it is slower than the method of QR decomposition with pivoting. On one hand, SVD is a good choice for high condition number problems that are often encountered in MS-CG applications. On the other hand, the results of singular value analysis identify problems of an ill-posed nature. Sparse solvers can be used with the block average algorithm in MS-CG calculations. In this case, the least-squares QR (LSQR) algorithm[42] is usually implemented since it is known to be a robust solver for ill-posed problems.

Various algorithm/solver strategies in the MSCGFM program are listed in Figure 4. For very complicated systems such as proteins with around 20 CG site types, the block average algorithm with sparse matrix format and LSQR solver is usually the only computationally affordable choice due to the very large memory requirement. The sequential accumulation algorithm is good at very ill-posed problems. When the matrix $\mathbf{F}$ is not very large and ill-conditioned, the normal equation algorithm is often a good choice since it is usually faster than the sequential accumulation algorithm. Both the normal equation and sequential accumulation algorithms involve intermediate dense matrices. Therefore, only dense solvers can be implemented in these cases, and the SVD solver is usually chosen.

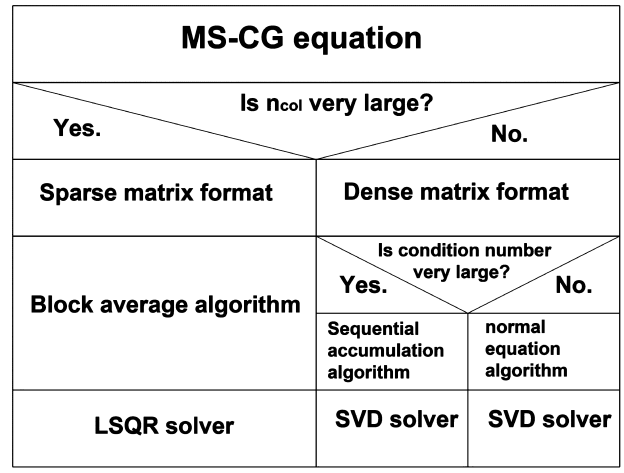


Figure 4. Various choices of MS-CG data structures, algorithms, and least-squares solvers.

4.4. Parallelization. If an efficient neighbor search algorithm is used in the construction of the $\mathbf{F}$ matrix, the most time-consuming part of the MS-CG calculation is the solution of the least-squares problem in eq 2a. Thus, parallel algorithms are applied only to the equation solving process. Since there are different equation solving algorithms for different types of MS-CG applications, separate parallelization strategies have to be developed for each case.

In the block average algorithm, the computation for each block is intrinsically independent. This makes the parallelization strategy quite straightforward. To parallelize the MS-CG calculation in this case, a number of independent processes are created, and each single process is responsible for the computation of a certain number of blocks. The average block solution from each process is then stored on hard disk, and the final average can be calculated from these solutions. No communication is necessary between any of the processes, so the parallel efficiency can be considered to be 100%. One disadvantage of this parallel strategy is that each process needs the same amount of memory as that for the serial computation. However, the block average algorithm is usually used for very large systems, and from Table 1 it can be seen that the matrix $\mathbf{F}$ is always very sparse in this case. As a result, the MS-CG calculation with sparse matrix format is not extremely demanding for memory, and the above strategy works in most applications. In principle, MPI-based sparse solvers may be implemented for problems with extremely large memory requirements that cannot be handled by the above strategy. However, the upper bound for the number of CG site types is around 20 to 30 for protein systems studied to date, and in practice these applications are tractable by the above parallelization strategy. In addition, the MPI-based sparse solver may reduce the parallel efficiency due to communication requirements. In the MSCGFM program, the parallelization for sparse algorithms is realized by the above strategy of independent processes, while the MPI-based algorithm is still under development.

The normal equation algorithm is parallelized by the same approach with independent processes. The MS-CG computation is divided into a number of processes for blocks of all-atom trajectory frames. The intermediate information that must be saved is block results for $\mathbf{F}^T\mathbf{F}$ and $\mathbf{F}^T\mathbf{f}$. The final results of $\mathbf{F}^T\mathbf{F}$ and $\mathbf{F}^T\mathbf{f}$ are then calculated on the basis of the block results from individual processes. Compared to the block average algorithm, the required hard disk space for intermediate results is much larger since both the left- and right-hand side information must be stored. This hard disk requirement is usually not a serious problem because the normal equation algorithm is used mostly for small systems.

In the sequential accumulation algorithm, the all-atom trajectory must be treated sequentially. Hence, it is not possible to divide the MS-CG calculation into independent processes and apply the above parallelization strategy. However, since the dense QR decomposition involved in the algorithm is included in MPI-based linear algebra software, such as SCALAPACK,[43] it is possible to implement MPI-based solvers in this case. Because of the need for interprocess communication, the parallel efficiency in this case may be much less than 100%, depending on the matrix structure and hardware/software configurations. However, since the sequential algorithm is usually used for small systems, in most applications no internode parallelization is necessary.

So far, the discussed parallelization strategies can be applied for both distributed and shared memory hardware. Presently, computer clusters with multi-CPU/multicore nodes and high speed internode connections are the standard in high performance scientific computing. Multicore CPUs have also become popular in workstations. In these cases, the computing node or CPU can be considered a computing unit with shared memory. Thus, it is desirable to implement parallelization for shared memory systems that benefits from the hardware structure. These parallelization approaches, such as the widely used OpenMP library,[44] can be used with the distributed memory parallelization discussed above or used alone for small- to medium-size problems. Fortunately, the Intel Math Kernel Library (MKL)[45] supports threaded LAPACK functions for linear algebra computations through the OpenMP environment. This is utilized for dense matrix algorithms in MS-CG calculations, and both the normal equation algorithm and the sequential accumulation algorithm are threaded.

4.5. Regularization. Since eq 2a is ill-possed in most medium- to large-size MS-CG applications, it is necessary to consider numerical regularization methods. For rank-deficient problems, simple regularization methods like the truncated SVD method that is implemented with the SVD solver in MS-CG calculations can generate reasonable results. However, more sophisticated regularization must be implemented for discrete ill-posed problems in MS-CG applications. For example, Liu et al.[46] successfully implemented a Bayesian regularization approach for the MS-CG problem. The results from Liu et al. show that the CG force curve in badly sampled interaction distances can be dramatically improved through the Bayesian treatment. However, the regularization method introduced here is different compared to that from Liu et al. in the following ways: First, the regularization from Liu et al. is designed for square equations; i.e., it is designed for the normal eq 5, whose matrix has a larger condition number than the matrix in eq 2a. The residual information that is important for regularization is also lost during the conversion to the normal equation. Second, the algorithm of Liu et al. involves a matrix inversion of $\mathbf{F}^T\mathbf{F}$. This makes the computation very expensive for large systems since the matrix inversion calculation is well-known to be slow.[47] Finally, the iterative scheme proposed by Liu et al. might converge slowly for complicated systems. The goal here is therefore to apply a regularization scheme that acts on the original eq 2a and is computationally inexpensive.

The Tikhonov regularization with L-curve criteria[38] for choosing regularization parameters is very successful for many ill-posed problems and is therefore appropriate to regularize eq 2a. In Tikhonov regularization, instead of solving eq 2a, one solves a regularized equation

$$\begin{bmatrix}\mathbf{F}\\ \lambda\mathbf{I}\end{bmatrix} \cong \begin{bmatrix}\mathbf{f}\\ 0\end{bmatrix} \tag{7}$$

Here, λ is a regularization parameter and $\mathbf{I}$ is the identity matrix. By adding the regularization term $\lambda\mathbf{I}$, the solution becomes smoother and the residual norm becomes larger. For the original eq 2a, a direct solution is often dominated by contributions from noise and round-off errors, and adding the regularization term can solve this problem at the cost of a slightly increased residual. Hence, the main objective in Tikhonov regularization is to choose a reasonable parameter λ in order to obtain the balance between solution accuracy and smoothness. This can be achieved by applying the L-curve criteria in which λ is chosen to give the "corner" of the residual norm-solution norm curve on a logarithm scale plot, where the corner is defined as the point with maximum curvature. In a standard L-curve procedure, a number of values of λ are chosen, and the corresponding residual norms and solutions norms are calculated. The results are then plotted on a log−log scale, and the corner point is found after numerical interpolation of the curve using splines. There are a few drawbacks, however, for applying this procedure in MS-CG calculations. First, the matrix $\mathbf{F}$ is not stored during the MS-CG calculation, and for each λ, it needs to be rebuilt by scanning the whole all-atom trajectory. Second, eq 7 must be solved for a number of different values of λ, which is the most expensive part of the L-curve calculation. Finally, the result of the corner point strongly depends on the interpolation scheme. Because calculating one point on the L-curve is expensive, usually only a few points on the plot can be obtained, and the interpolation accuracy can be difficult to guarantee.

Fortunately, if SVD can be performed for $\mathbf{F}$, the curvature for a point on the L-curve can be expressed analytically as a function of λ and singular values and vectors.[48,49] In this way, the procedure to find the corner turns out to be a one-dimensional minimization problem involving λ. Therefore, the atomistic trajectory needs to be read only once, and only one SVD needs to be performed for the original matrix $\mathbf{F}$. This seems to be much more efficient than the above approach of computing each L-curve point. However, the

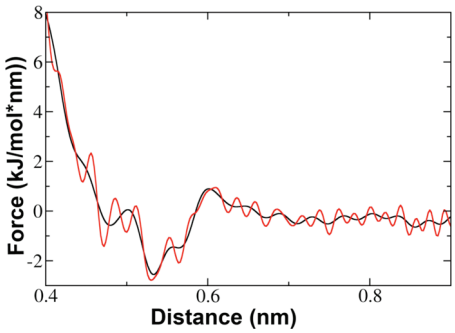


Figure 5. The CG forces for water−water interactions. The red line is the original MS-CG force, and the black line is the force after regularization. Only force values in the range of 0.4 to 0.9 nm are plotted to clearly show the overfitting phenomenon.

huge size of **F** makes the direct SVD prohibitive in most cases. Recall that in the sequential accumulation algorithm the matrix **R** in eq 6 has the same singular values as **F**, and eq 6 has the same solution and residual as the original eq 2a. Thus, regularization can be applied to eq 6 instead as

$$\begin{bmatrix} \mathbf{R} \\ 0 \\ \lambda\mathbf{I} \end{bmatrix} \cong \begin{bmatrix} \mathbf{d} \\ e \\ 0 \end{bmatrix} \tag{8}$$

in which SVD of **R** is usually affordable.[37] For extremely large systems that are usually handled by the block average algorithm with a sparse solver, it is still possible to convert the L-curve corner-finding procedure to a one-dimensional minimization problem involving λ. However, a least-squares equation with a size similar to eq 2a needs to be solved for each λ, and the whole λ optimization case is much more expensive than the approach with SVD of eq 6.[48]

A numerical example of a system with 1000 TIP3P[50] water molecules is shown in Figure 5. In this example, a B-spline basis function with the order $k = 40$ and spacing 0.01 nm is applied. An unreasonably high order B-spline is used to demonstrate the possibility of the overfitting phenomenon in MS-CG calculations. Since the matrix for this application is small, it is possible to do SVD for the left-hand side matrix in eq 6. Therefore, the curvature of the L-curve can be calculated analytically, and the resultant one-dimensional optimization value for λ is 1.5849. The black line in Figure 5 is the CG force curve after regularization with the λ value. It is shown in Figure 5 that the original CG force curve has many unphysical fluctuations corresponding to overfitting. However, the force curve is smoothed after regularization, and these artificial fluctuations disappear. Since the statistical sampling for this CG water system is reasonable, it is hard to observe the effect of regularization until a very large k is used, and no regularization is needed for a normal k value. It is expected that for other more ill-posed MS-CG problems the difference between the results with and without regularization appears to be more obvious, and regularization becomes necessary.

5. Miscellaneous Algorithms

In MS-CG calculations, most basis functions in eq 3 are spline functions. In order to define a spline interpolation, the minimum and maximum distances for each interaction need to be determined. A complete search for these distance ranges is therefore performed before defining spline basis sets. There are two advantages for doing the distance search rather than guessing the ranges arbitrarily. On one hand, this search assures minimal memory spaces are allocated for basis functions. Spline basis functions are assigned to sampled distances only after the search. On the other hand, excluding basis functions corresponding to unsampled distances reduces the condition number of the matrix **F** as pointed out by Noid et al.[23] For large systems, the distance searching can be time-consuming, though it is still much faster than solving the least-squares equation. Since only minimum and maximum distances for each interaction are needed, the distance searching can be divided into independent processes for small trajectory blocks, which is straightforward to be parallelized for large systems. Sometimes RDFs or bond/angle/dihedral distributions are calculated for other purposes. In this case, distance ranges can be extracted from distribution function data, and the distance search is no longer necessary.

In the MS-CG application to ionic liquids by Wang et al.,[24] the bonded interactions are determined from an inverse Boltzmann approach because the bonded parameters are usually more affected by statistical noise. CG parameters not easily obtained from the MS-CG scheme can typically be determined by inverting distribution functions or using some knowledge based approach. The iterative procedure suggested by Wang et al.[24] can be applied to determine those CG parameters. The MSCGFM code allows an arbitrary set of the ϕ coefficients to be chosen in advance, input into the calculation, and held fixed in the variational calculation. This makes it possible to have the MS-CG variational principle determine only a few critical CG interactions for complex systems, and the computation expense can be reduced. The variational calculation with some coefficients held fixed is of the same form as eq 2a, with a new **F** matrix with fewer columns than the original matrix, a new ϕ matrix that includes the parameters that are to be varied, and a new **f** vector with the same length as the original vector.

One advantage of specifying some interaction parameters in advance and keeping them fixed in the variational calculation is that it speeds up the MS-CG calculation, especially when a sparse solver is used. In most biomolecular all-atom systems, the number of solvent molecules is very large, and in the corresponding CG system most CG sites are solvent sites. Consequently, most elements in **F** are determined by forces acting on the solvent in the atomistic simulations. If the solvent−solvent interaction parameters are not included in the variational calculation, the new **F** matrix has many fewer nonzero elements than the original **F**, and the MS-CG calculation is much less expensive for a sparse solver. Usually the solvent−solvent interaction is obtained from a separate MS-CG calculation for a pure solvent system. For dense solver cases, removing solvent−solvent interactions from the variational calculation does not change the computation time dramatically since the matrix dimension is only changed slightly.

Table 2. MS-CG Calculation Benchmarks for Selected Molecular Systems[a]

system	basis function	CG type	speed (frames/hour)
Water	linear spline	1	4.99×10^4
DPPC	linear spline	7	4.32×10^2
DPPC	B-spline	7	3.26×10^3
T4 lysozyme	B-spline	32	19.5

[a] Details of each system are described in the main text.

6. Benchmarks

Three systems were chosen to generate benchmarks of the MSCGFM program. The first system contained 1000 TIP3P water molecules,[50] and a CG mapping was applied to this system to obtain a one-site CG water model. For this system, a linear spline of a grid spacing of 0.005 nm was applied. The second system contained 64 DPPC lipids and 3846 TIP3P water molecules.[50] A solvent-free DPPC lipid CG model was then generated from the all-atom systems, and the lipid CG mapping was similar to that used by Izvekov and Voth.[26] As a result, there were 960 CG sites and 7 CG site types in the CG system. Bonds and angles were treated as 1−2 and 1−3 two-body interactions, respectively, and the CG potential contains no explicit dihedral terms.[26] Two different types of basis sets were applied to this system. The first type were linear spline basis functions with a spacing of 0.01 nm for the nonbonded interactions and 0.005 nm for the bonded interactions. The second type were B-spline basis functions with a spacing of 0.04/0.01 nm and B-spline order k as 6/4 for the nonbonded/bonded interactions. In practice, these B-spline parameters give a CG force field result that is very similar to that from linear splines. The normal equation algorithm and SVD dense solver were applied for the above two systems to solve the least-squares problem. The third system contained a bacteriophage T4 lysozyme and 9733 explicit water molecules. After the all-atom to CG mapping there were 9771 CG sites and 32 CG site types in the CG system. All adjacent protein CG sites were connected by bonds, which made the total number of bonds 29. The B-spline basis functions for this system had a spacing of 0.03 nm for the nonbonded interactions and 0.01 nm for the bonded interactions. The B-spline order was 4 for all interactions. This application was to show the program's ability to treat very large systems with the water−water interaction inputted from a table for computational efficiency. The MS-CG calculation for this protein system was performed with the block average algorithm and LSQR sparse solver.

Table 2 shows the computation speed for each MS-CG calculation for the three tested systems. All computations in Table 2 were performed on a workstation with Intel dual core Xeon 2.0 GHz CPUs. Only one core of the CPUs was used for each calculation, and the Intel MKL package[45] was used for dense matrix computations. The LSQR computation was performed using a compiled FORTRAN 77 code. Computational speed is defined by the number of all-atom configurations (frames) that are processed per hour. It is seen in Table 2 that MS-CG calculations for systems with less than 10 CG site types are usually very efficient on a single processor. Note that for most MS-CG calculations at least several thousand all-atom configurations are necessary. For better statistical sampling and smoother output force curves, tens of thousands of atomistic configurations are desirable depending on the system complexity. On the basis of the speed data in Table 2, these requirements can be easily achieved for small systems. However, the matrix column dimension N_D is proportional to the square of the number of CG site types, which causes the required computer time to increase dramatically when the number of CG site types becomes large. For example, the third benchmark system has 32 CG site types, and it is extremely hard to solve this MS-CG problem using dense matrix algorithms considering the memory and computer time required. For this large system, the required computer time even prohibits calculation with a sparse algorithm. Therefore, the water−water interaction is inputted from an external table, which dramatically speeds up the calculation. The speed results show it is able to perform the MS-CG calculation for this complex system using tens of CPUs and a few days. Thus, this application is a good example to demonstrate MSCGFM's ability to treat complex systems.

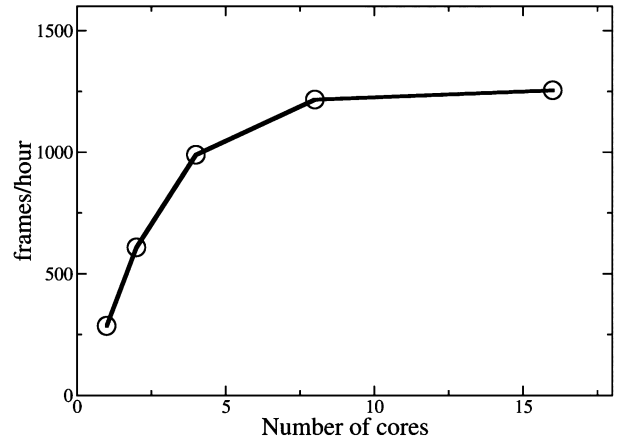


Figure 6. Scaling of an OpenMP parallelized MS-CG calculation for a DPPC lipid CG system.

From Table 2, it is also seen that for small- to medium-sized MS-CG systems with around 10 CG site types, applying B-spline basis functions can greatly increase the computation efficiency. This is because the matrix column dimension is reduced by using a smaller number of basis functions. For very large problems with a sparse algorithm, the computation speed depends on both matrix dimension and nonzero element percentage. Using higher-order basis functions usually makes the matrix **F** less sparse. Therefore, there is no straightforward relation between the type of basis functions and the computation efficiency if a sparse algorithm is used.

Figure 6 shows the scaling for dense normal equation algorithms with OpenMP support. The system is the DPPC system described earlier with linear spline basis functions. These calculations were performed on 16-core computing nodes on an AMD Opteron cluster with OpenMP supported Intel MKL software. It is clearly seen that the calculation scales well up until 8 cores. Considering that computing nodes on most high performance computing clusters are comprised of 2 to 16 cores, dense matrix MS-CG calculations can be well parallelized for these hardware structures.

7. Conclusions

A number of algorithms for MS-CG calculations have been introduced in the recently developed MSCGFM program. The MS-CG approach has been demonstrated to be a successful method to generate CG force field parameters from all-atom trajectories. However, computational efficiency and stability are critical issues when applying the MS-CG method to complex systems. In order to speed up the calculation, various dense and sparse matrix algorithms, which are appropriate for different types of CG systems, are incorporated in the MSCGFM code. In particular, the implementation of sparse algorithms enables the MS-CG program to deal with complicated biomolecular systems, such as proteins. Tikhonov regularization is also proposed to regularize ill-posed MS-CG problems for certain CG systems. Figure 4 shows a flowchart for the implementation of the code for a given MS-CG calculation.

Future work will involve the development of better parallelization approaches for the MS-CG calculations on distributed memory machines, especially for the sparse matrix algorithms. New basis functions corresponding to various physical interactions will also be incorporated.

The code of MSCGFM is available on request and will soon be released under a public license.

Acknowledgment. This research was supported by a Collaborative Research in Chemistry grant from the National Science Foundation (CHE-0628257). Computer resources were provided by the National Science Foundation through TeraGrid computing resources administered by the Pittsburgh Supercomputing Center, the San Diego Supercomputer Center, the National Center for Supercomputing Applications, the Texas Advanced Computing Center, and Argonne National Laboratories. The authors gratefully acknowledge Dr. Edward Lyman, Dr. Ron Hills and Dr. Sven Jakobtorweihen for critical reading of the manuscript.

References

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

(2) MacKerel, Jr., A. D.; Brooks III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*; John Wiley & Sons: Chichester, 1998; Vol. 1; pp 271.

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

(4) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(5) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205.

(6) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

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

(8) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2008.

(9) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812.

(10) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819.

(11) Lyubartsev, A. P.; Laaksonen, A. *Phys. Rev. E* **1995**, *52*, 3730.

(12) Reith, D.; Putz, M.; Muller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624.

(13) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869.

(14) Henderson, R. L. *Phys. Lett. A* **1974**, *49*, 197.

(15) Chayes, J. T.; Chayes, L.; Lieb, E. H. *Commum. Math. Phys.* **1984**, *93*, 57.

(16) Chayes, J. T.; Chayes, L. *J. Stat. Phys.* **1984**, *36*, 471.

(17) Lyubartsev, A. P.; Laaksonen, A. *Phys. Rev. E* **1997**, *55*, 5689.

(18) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.

(19) Johnson, M. E.; Head-Gordon, T.; Louis, A. A. *J. Chem. Phys.* **2007**, *126*, 144509.

(20) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469.

(21) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.

(22) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(23) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.

(24) Wang, Y. T.; Izvekov, S.; Yan, T. Y.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 3564.

(25) Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 11566.

(26) Izvekov, S.; Voth, G. A. *J. Chem. Theory Comput.* **2006**, *2*, 637.

(27) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289.

(28) Thorpe, I. F.; Zhou, J.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 13079.

(29) Lu, L. Y.; Voth, G. A. *J. Phys. Chem. B* **2009**, *113*, 1501.

(30) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2009**, *113*, 4443.

(31) Das, A.; Andersen, H. C. *J. Chem. Phys.* **2009**, *131*, 034102.

(32) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles*; Taylor & Francis, Inc.: Bristol, PA, USA 1988.

(33) Quentrec, B.; Brot, C. *J. Comput. Phys.* **1975**, *13*, 430.

(34) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, L. S.; Demmel, J.; Dongarra, J. J.; Croz, J. D.; Hammarling, S.; Greenbaum, A.; McKenney, A.; Sorensen, D. *LAPACK Users' guide (third ed.)*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1999.

(35) Saad, Y. *Iterative Methods for Sparse Linear Systems*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2003.

(36) Antonakos, J. L.; Mansfield, K. C. *Practical Data Structures Using C/C++ with 3.5 Disk*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 1998.

(37) Lawson, C. L.; Hanson, R. J. *Solving Least Squares Problems*; Prentice-Hall, Englewood Cliffs, NJ, 1974.

(38) Hansen, P. C. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1998.

(39) Golub, G. H.; Van Loan, C. F. *Matrix computations* (3rd ed.); Johns Hopkins University Press: Baltimore, MD, USA 1996.

(40) Hansen, P. C. *BIT* **1987**, *27*, 534.

(41) Savelyev, A.; Papoian, G. A. *J. Phys. Chem. B* **2009**, *113*, 7785.

(42) Paige, C. C.; Saunders, M. A. *ACM Trans. Math. Softw.* **1982**, *8*, 195.

(43) Blackford, L. S.; Choi, J.; Cleary, A.; D'Azeuedo, E.; Demmel, J.; Dhillon, I.; Hammarling, S.; Henry, G.; Petitet, A.; Stanley, K.; Walker, D.; Whaley, R. C. *ScaLAPACK user's guide*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1997.

(44) Chandra, R.; Dagum, L.; Kohr, D.; Maydan, D.; McDonald, J.; Menon, R. *Parallel programming in OpenMP*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001.

(45) Intel Math Kernel Library for Linux OS User's Guide. http://www.intel.com/software/products/ (accessed Jan 11, 2010).

(46) Liu, P.; Shi, Q.; Hal Daumé, H., III; Voth, G. A. *J. Chem. Phys.* **2008**, *129*, 214114.

(47) Press, W. H.; Teukolsky, S. A.; Vettering, W. T.; Flannery, B. P. *Numerical Recipes in C, The Art of Scientific Computing, Second Edition*; Cambridge University Press: New York, NY, USA, 1997.

(48) Hansen, P. C. "The L-curve and Its Use in the Numerical Treatment of Inverse Problems"; in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering, 2000.

(49) Hansen, P. C.; O'Leary, D. P. *SIAM J. Sci. Comput.* **1993**, *14*, 1487.

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

CT900643R

JCTC Journal of Chemical Theory and Computation

Membrane/Toxin Interaction Energetics via Serial Multiscale Molecular Dynamics Simulations

Chze Ling Wee, Martin B. Ulmschneider,† and Mark S. P. Sansom*

Department of Biochemistry and Oxford Centre for Integrative Systems Biology, University of Oxford, South Parks Road, Oxford, OX1 3QU, United Kingdom

Received December 4, 2009

Abstract: Computing free energies of complex biomolecular systems via atomistic (AT) molecular dynamics (MD) simulations remains a challenge due to the need for adequate sampling and convergence. Recent coarse-grained (CG) methodology allows simulations of significantly larger systems ($\sim 10^6$ to 10^8 atoms) over longer (μs/ms) time scales. Such CG models appear to be capable of making semiquantitative predictions. However, their ability to reproduce accurate thermodynamic quantities remains uncertain. We have recently used CG MD simulations to compute the potential of mean force (PMF) or free energy profile of a small peptide toxin interacting with a lipid bilayer along a 1D reaction coordinate. The toxin studied was VSTx1 (Voltage Sensor Toxin 1) from spider venom which inhibits the archeabacterial voltage-gated potassium (Kv) channel KvAP by binding to the voltage-sensor (VS) domains. Here, we re-estimate this PMF profile using (i) AT MD simulations with explicit membrane and solvent and (ii) an implicit membrane and solvent (generalized Born; GBIM) model where only the peptide was explicit. We used the CG MD free energy simulations to guide the setup of the corresponding AT MD simulations. The aim was to avoid local minima in the AT simulations which would be difficult over shorter AT time scales. A cross-comparison of the PMF profiles revealed a conserved topology, although there were differences in the magnitude of the free energies. The CG and AT simulations predicted a membrane/water interface free energy well of −27 and −23 kcal/mol, respectively (with respect to water). The GBIM model, however, gave a reduced interfacial free energy well (−12 kcal/mol). In addition, the CG and GBIM models predicted a free energy barrier of +61 and +96 kcal/mol, respectively, for positioning the toxin at the center of the bilayer, which was considerably smaller in the AT simulations (+26 kcal/mol). Thus, we present a framework for serially combining CG and AT simulations to estimate the free energy of peptide/membrane interactions. Such approaches for combining simulations at different levels of granularity will become increasingly important in future studies of complex membrane/protein systems.

Introduction

Accurate determination of free energies is crucial for the understanding of various biophysical systems. Molecular dynamics (MD) simulations have been used to estimate free energies, e.g., by counting the number of events along the reaction coordinate(s) of interest.[1] In the context of membrane proteins and channels, for example, MD simulations have been used to calculate the free energy of ion conduction through ion channels[2,3] and related pores,[4,5] and the free energy of partitioning of amino acid side chain analogues into lipid bilayers.[6] These calculations are founded on the basis that well sampled distributions can be obtained from

* To whom correspondence should be addressed. Phone: +44 1865 613306. Fax: +44 1865 613238. E-mail: mark.sansom@bioch.ox.ac.uk.

† Current address: Department of Physiology and Biophysics, University of California at Irvine, Irvine, CA 92697-4560.

10.1021/ct900652s
Published on Web 02/04/2010

finite-time simulations. Atomistic (AT) MD simulations can routinely address systems of $\sim 10^5$ to 10^6 atoms over $\sim 10^2$ to 10^3 ns. This is likely to be insufficient for sampling all but the simplest systems.[1] For example, many biophysical processes of interest (e.g., ion channel gating) occur over longer (e.g., >1 μs) time scales. Thus, achieving adequate sampling in free energy simulations of membrane/protein systems remains challenging.

Recent coarse-grained (CG) simulation methodology has generated considerable interest as it permits simulations of larger systems over longer time scales.[7−15] In the context of membrane/protein systems, CG simulations have been used to predict the orientation of proteins in lipid bilayers,[16−20] the effect of hydrophobic mismatch on membrane/protein dynamics,[20−23] protein conformational changes,[24,25] and protein/protein interactions.[15,26] Thus, for certain applications, CG simulations can provide insights into systems and events that are out of reach of AT simulations.[8,27] CG simulations have also been used to perform more rigorous calculations such as, e.g., estimation of free energies of peptide/membrane interactions.[20,28−30] Although CG simulations can result in well sampled distributions, their ability to reproduce thermodynamic properties needs to be better evaluated.[31,32]

More recently, there has been interest in combining simulations performed at different levels of granularity.[33−36] One approach is to do this serially[34] where simulations at one level of granularity can be used to generate starting coordinates for simulations at a different level of granularity.[19,37] Simulations at one level of granularity can also be used to parametrize simulations at a different level of granularity.[33] Alternatively, one could do this in parallel where, e.g., a region of interest is described atomistically with the remainder of the system described with a CG model, with special treatments at the AT/CG boundary.[14,35,38]

We have recently used CG MD simulations to calculate the 1D potential of mean force (PMF) or free energy profile of a small peptide toxin interacting with a lipid bilayer via an umbrella sampling protocol.[20] The reaction coordinate corresponds to the position, projected along the bilayer normal (z axis), of the center of mass (com) of the toxin with respect to the com of the membrane. The toxin studied was VSTx1 (Voltage Sensor Toxin 1) from spider venom which inhibits the archeabacterial voltage-gated potassium (Kv) channel KvAP by interacting with the voltage-sensors.[39−42] Thus, VSTx1 inhibits KvAP by altering the energetics of voltage-dependent gating.[41,42] VSTx1 is a small globular protein (34 residues) with a distinct amphipathic molecular surface,[41−45] with one half of the toxin predominantly hydrophobic and the other half predominantly polar (Figure 1A). VSTx1 is positively charged (+3) at pH 7. It is structurally stable due to presence of 3 internal disulfide bridges.[46] Gating-modifier toxins such as VSTx1 gain access to the VS by first binding to the membrane/water interface.[41,42,44,47,48] Membrane partitioning is consistent with our CG free energy profiles which revealed a location at the membrane/water interface when VSTx1 interacted with lipid bilayers.[20] From a biophysical perspective, gating-modifier toxins such as VSTx1 have proved to be valuable characterization tools for studying voltage-gated ion channels.[41,42] From a computational point of view, the interaction of VSTx1 with lipid bilayers has been studied via CG[20] MD and via AT[49] MD simulations and is a tractable and well understood problem. It is therefore an ideal test system for a multiscale approach for computing the free energy of peptide/membrane interactions.

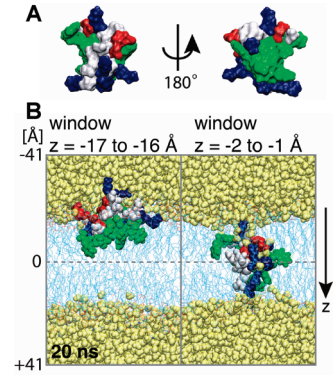


Figure 1. (A) VSTx1 has an amphipathic molecular surface with a hydrophilic (left) and a hydrophobic (right) side. Basic, acidic, polar, and hydrophobic residues of VSTx1 are colored blue, red, white, and green, respectively. (B) Atomistic (AT) umbrella sampling simulations were used to compute a 1D PMF profile where the reaction coordinate corresponds to the position, projected along the bilayer normal (z axis), of the center of mass (com) of the toxin with respect to the com of the membrane (the bilayer center is at $z \sim 0$ Å). A total of 82 umbrella sampling simulations or "windows" spaced 1 Å apart were used to sample from the extracellular (EC) solvent, across the bilayer, and into the intracellular (IC) solvent ($z = -41$ to $+41$ Å). Snapshots were taken at 20 ns for toxin locations at the free energy well at the bilayer/water interface ($z = -17$ to -16 Å; left) and in the hydrophobic core of the bilayer ($z = -2$ to -1 Å; right). POPC carbon, oxygen, nitrogen, and phosphorus atoms are colored cyan, red, blue, and gold, respectively, in a lines representation. Waters are colored yellow.

Here, we re-estimated the 1D PMF profile of VSTx1 interacting with a pure palmitoyl−oleoyl−phosphatidyl−choline (POPC) bilayer with *AT* simulations. An umbrella sampling protocol was used, and the reaction coordinate corresponds to the position, projected along the bilayer normal, of the com of the toxin with respect to the com of the membrane. The availability of considerable CG simulation data on this system (>30 μs[20]) permits a multiscale approach. We used the CG simulations to guide the setup of the AT simulations. Our motivation was to avoid local minima in the AT simulations, important for meaningful free energy estimates (see Theory and Methods), which is difficult to achieve over shorter AT time scales. As a further test, we used an implicit membrane/implicit solvent model[50] based on the generalized Born framework[51,52] (GBIM) to derive an estimate of a PMF profile where only the toxin was explicit. Cross-comparisons of the PMF profiles (AT, CG, and GBIM) revealed differences in the magnitude of the free energies although the overall topology of the landscapes was preserved. The PMF profiles reinforce the view whereby VSTx1 partitions from water into the membrane/water interface, with a considerable free energy barrier for positioning the toxin at the center of the membrane. By using

serial multiscale MD simulations, we have obtained a unified view of how VSTx1 interacts with a membrane.

Theory and Methods

Umbrella Sampling MD Simulations to Derive a Potential of Mean Force Profile. The PMF $W(\varepsilon)$ along a reaction coordinate ε is the constrained free energy and can be defined from the average distribution $\langle\rho(\varepsilon)\rangle$

$$W(\varepsilon) = W(\varepsilon^*) - k_B T \ln\left[\frac{\langle\rho(\varepsilon)\rangle}{\langle\rho(\varepsilon^*)\rangle}\right]$$

where ε^* and $W(\varepsilon^*)$ are arbitrary constants.[53] The average distribution function along the coordinate ε is obtained from a Boltzmann weighted average

$$\langle\rho(\varepsilon)\rangle = \frac{\int \delta(\varepsilon'[R] - \varepsilon) e^{-U(R)/k_B T}\, dR}{\int e^{-U(R)/k_B T}\, dR}$$

where $U(R)$ represents the total energy of the system as a function of the coordinates R, and $\varepsilon'[R]$ is a function depending on one or more degrees of freedom in the system.[53]

MD simulations can be used to sample $\langle\rho(\varepsilon)\rangle$. However, the presence of energy barriers along ε, together with simulations of finite duration, are likely to prevent adequate sampling of $\langle\rho(\varepsilon)\rangle$. With umbrella sampling MD simulations, one performs an ensemble of N biased simulations or "windows" where an external biasing potential $\omega(\varepsilon)$ is applied to force the system to sample over the range of interest along ε. Each window samples the neighborhood of a chosen value of ε, and one obtains N biased distributions which need to be unbiased and combined to obtain a single unbiased distribution.[53,54] In the current study, we used the weighted histogram analysis method (WHAM).[55] We emphasize that the integral in the average distribution function implies that all nonsampled degrees of freedom "perpendicular" to the reaction coordinate ε must equilibrate in *all* windows before collecting distributions along ε for the PMF to be meaningful.

Serial Multiscale CG and AT Simulations. MD simulations[18,20,49] and experiments[41,56] suggest VSTx1 and related gating-modifier toxins have a distinct orientation when bound to the lipid bilayer. At its optimal location of interaction at the membrane/water interface, the polar residues of VSTx1 are directed toward solvent where they interact with the lipid headgroups, with the hydrophobic residues exposed to the lipid tails (Figure 1B; left panel).[20] To obtain a meaningful PMF profile, it is crucial that the orientation of VSTx1 equilibrates at each point along the reaction coordinate z (the z axis corresponds to the bilayer normal). This was achieved in our recent CG free energy simulations.[20] We had used 104 independent CG windows (each of duration 40 ns) spaced 1 Å apart along z to sample configurations of VSTx1 from the extracellular (EC) solvent, across a POPC bilayer, and into the intracellular (IC) solvent. An *identical* initial orientation of the toxin relative to the bilayer was used in all windows (as shown in Figure 2A; 0 ns). We quantified the orientation of VSTx1 by the angle of its hydrophobic

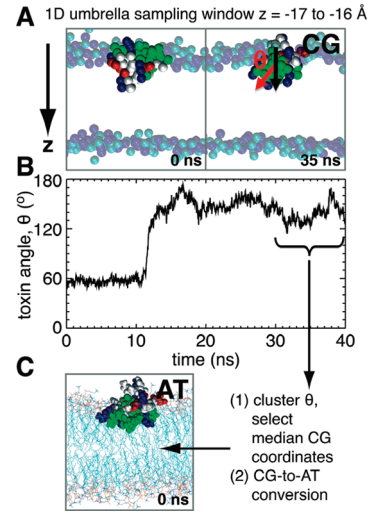


Figure 2. Serial multiscale CG[20] and AT umbrella sampling simulations. (A, B) An identical initial orientation of VSTx1 (relative to the membrane) was previously used for all CG windows (window z = 16 to 17 Å shown; each CG window was simulated for 40 ns).[20] The orientation of VSTx1 (as defined by the angle of the hydrophobic moment of the toxin[57] θ) relative to the bilayer normal equilibrated by 20 ns in all windows. (C) Each CG window was clustered on the backbone of VSTx1 and over the final 10 ns (i.e., 30 to 40 ns/window; after toxin orientation had equilibrated). The median of the cluster, which corresponds to the system with the most frequently observed equilibrated orientation of VSTx1, was used to guide the setup of the corresponding AT window. POPC phosphates and cholines are colored blue and cyan, respectively, in the CG snapshots. All other CG particles were not shown for clarity.

moment (defined as the vector sum of the hydrophobicity of each constituent residue of a peptide[57]) relative to the bilayer normal (θ; Figure 2A; note that θ ranges from 0° to 180°). θ equilibrated by 20 ns for CG windows where VSTx1 interacted with the bilayer to effectively minimize the hydrophobic mismatch between the toxin and its environment (Figure 2B). Thus, a distinct toxin orientation as a function of z exists. The only exceptions were windows where the toxin was located in water, away from the bilayer. Here, VSTx1 tumbled randomly in water.

To set up the AT umbrella sampling simulations, we clustered each CG window on the backbone of VSTx1 and over the final 10 ns (i.e., 30 to 40 ns/window; after toxin orientation had equilibrated) using the full linkage algorithm.[58] The median of the cluster, which corresponds to the system with the most frequently observed equilibrated orientation of VSTx1, was selected. The NMR structure of VSTx1[44] was least-squared-fitted on the CG toxin coordinates, and the AT toxin was subsequently modeled into an AT POPC bilayer (Figure 2C; see Umbrella Sampling MD Simulation Setup).

Thus, unlike the CG free energy simulations where an arbitrary initial orientation of VSTx1 was used in each window,[20] knowledge gained from the CG simulations was used to initiate the orientation of the toxin in the AT free energy simulations. We therefore assumed (i) the orientation of VSTx1 relative to the bilayer was the overwhelming factor

that would have influenced the accuracy of a 1D PMF along z, and (ii) the CG model[59] was able to predict θ accurately. The latter is reasonable as, e.g., the CG model has successfully predicted the insertion and orientation of a range of membrane/membrane-associated proteins in lipid bilayers[16−18] including VSTx1.[20]

Umbrella Sampling MD Simulation Setup. To perform AT umbrella sampling, we used 82 windows along z spaced 1 Å apart ($z = -41$ to 41 Å; the bilayer center is at $z \sim 0$ Å; Figure 1B) to sample from the EC solvent, across the bilayer, and into the IC solvent. For each window, 20 ns AT MD simulation was employed. In total, we accumulated >1.6 μs of AT MD simulation data. The z coordinate of the com of VSTx1 was restrained relative to the z coordinate of the com of the entire bilayer with a harmonic biasing potential (e.g., in window $z = -41$ to -40 Å, the com of VSTx1 was restrained at -40.5 Å relative to the com of the bilayer). We used a pre-equilibrated POPC bilayer containing 128 lipids (courtesy of Peter Tieleman; moose.bio.ucalgary.ca). To generate initial coordinates for VSTx1 in the bilayer, we used a protocol similar to that employed in previous simulations of VSTx1[20,49] and related toxins.[18] Briefly, we modeled VSTx1 in the bilayer by gradually scaling the coordinates of the toxin to its full size in 10 steps, with 100 steps of steepest-descent energy minimization at each step to allow the lipids to adjust their conformations to host the toxin. The intramolecular interactions (i.e., interactions between the atoms of VSTx1) were switched off; however, interactions between VSTx1 and the lipids remained. The lipid coordinates from each previous step were kept while the original toxin coordinates were rescaled. At each step, any lipid whose phosphorus atom was within 2.5 Å of any toxin atom was removed. A cutoff of 2.5 Å was found to be optimal for removing steric clashes between VSTx1 and the lipids, while ensuring the lipids were adequately packed around the toxin. Across all windows, no more than 7 lipids/window were removed.

We had previously constrained our CG free energy simulations of VSTx1 interacting with a membrane by limiting excessive membrane deformation, to study the effects of membrane deformability on free energies.[20] We used a similar setup in our AT free energy simulations. Thus, positional restraints where applied on the phosphorus atoms of all lipids along the bilayer normal z.

AT umbrella sampling MD simulations were performed using GROMACS 3.2.1 (www.gromacs.org)[60,61] using the GROMOS-96 force-field[62] and Berger parameters for POPC lipids.[63,64] VSTx1 was kept in the default protonation state for pH 7. Each system was solvated with SPC waters.[65] Cl^- counterions were added to keep each system electrically neutral. Long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) method,[66] employing a grid spacing of ~ 1 Å^{-1} and an interpolation order of 4. A cutoff of 12 Å was used for the real space portion of the Ewald sum and the Lennard-Jones interactions. The LINCS algorithm[67] was applied to constrain all covalent bonds, and the SETTLE algorithm[68] was used to maintain the geometry of the water molecules. Each system was temperature coupled with a Berendsen thermostat[69] with a weak coupling constant of 0.1 ps and a reference temperature of 310 K. Semi-isotropic pressure coupling with a Berendsen barostat in x and y at 1 bar with a coupling constant of 1.0 ps and a compressibility value of 4.6×10^{-5} bar^{-1} was used. Prior to production, a short MD simulation of duration 0.5 ns/window (with positional restraints applied on the toxin, and on the phosphorus atoms of the lipids in z only) was performed to allow the lipids to further repack around the toxin, and to allow the waters and counterions to settle. This was followed by a production simulation of duration 20 ns/window. The time step of integration was 2 fs. Positional restraints and the biasing potential utilized a force constant of 10 kJ mol^{-1} Å^{-2}. All simulations utilized a TM voltage of 0 V. The biased probability distributions were combined and unbiased with an implementation of WHAM[55] (courtesy of Alan Grossfield; membrane.urmc.rochester.edu), using 100 bins over -41 Å and $+41$ Å and a tolerance of 0.0001.

The Generalized Born Membrane. The generalized Born implicit membrane (GBIM) used here has been described in detail in a previous publication.[70] In brief, the membrane is modeled as a low dielectric zone in a uniform aqueous implicit solvent (with a dielectric constant $\epsilon_{water} = 80$). The zone has a Gaussian cross-section becoming increasingly inaccessible to the solvent toward its center. Both the protein interior and the membrane are assumed to have the same interior dielectric constant of $\epsilon_{membrane} = 2$. The Born radii were calculated using the fast asymptotic pairwise summation, using the original OPLS all atom parametrization by Qiu and Still.[71] This yields excellent results in predicting experimental free energies of solvation as well as hydration effects on conformational equilibria.[72] The membrane was introduced by modifying the pairwise summation to solute atoms, the self-solvation terms $\Gamma(z_i, L)$, and the atomic volumes $V(z_i)$, which were made to vary smoothly between full solvation and a limiting value for burial at the center of the membrane. A Gaussian shape

$$\Gamma(z_i) = g_{bulk} + (g_{centre} - g_{bulk})\, e^{\gamma(z_i^2/L^2)}$$

was used, where g_{bulk} is the limiting value of Γ at a large distance from the membrane (i.e., $z \gg L$) corresponding to the self-solvation term of the unmodified generalized Born method, while g_{center} is the value of Γ at the membrane center. We used a Gaussian with $\gamma = -3.0$ and membrane half width of $L = 15$ Å, which corresponds roughly to the hydrophobic profile of a dipalmitoyl−phosphatidyl−choline (DPPC) lipid bilayer. A value of $g_{center} = -7.67$ kcal/mol was used, as reported previously.[73,74] Gaussians were chosen in good agreement with experimental evidence from lipid distortion[75,76] and X-ray and neutron diffraction experiments on fluid liquid-crystalline bilayers.[77] The OPLS-AA force field[78] was used to describe the toxin.

The nonpolar part of the solvation free energy is modeled using an effective surface tension associated with the solvent accessible surface area (SA).[71] As it is moved toward the center of the membrane, the surface energy contribution of each atom is scaled down by an exponential switch at the membrane interface ($z = \pm 15$ Å). Thus, for distances far from the membrane (i.e., $z \gg L$), the nonpolar contribution

is included with the positive surface tension of solvation in water, while in the center of the membrane the surface tension is negative (i.e., energy is gained by moving into the membrane from the gas phase), as determined experimentally.[79]

Calculating the Minimal Energy Conformation to Derive a Potential of Mean Force Profile. The minimal energy configuration of VSTx1 in a GBIM membrane was calculated by exploring the entire translational and rotational space of VSTx1 in the membrane, treating the toxin as a rigid body. Previous AT simulations of VSTx1 in water and lipid bilayers[49] revealed the toxin to be conformationally stable, with little structural drift from the initial toxin conformation over multinanosecond time scales, and no unfolding. The principal axis of the toxin was determined through diagonalization of the inertia tensor using only the heavy backbone atoms. The tilt angle was defined as the angle of the principal axis with respect to the membrane normal, while the rotation angle was defined as the angle of rotation around the principal axis.

The toxin was translated from $z = -50$ Å to $+50$ Å along the membrane normal (membrane center = 0 Å) in 0.5 Å steps. At each step, the toxin was rotated through all space to find the orientation of minimum energy by first tilting it with respect to the membrane normal and subsequent rotation around its principal axis until all tilt and rotational states have been sampled with a step size of 1°. The lowest energy conformation encountered was then subjected to a rigid body minimization in order to locate the precise location of the global energy minimum. Because the membrane and solvent are implicit (i.e., always at equilibrium) and VSTx1 is treated as a rigid body (which is not unreasonable as the toxin is conformationally stable), the potential energies approximate to free energies. Thus, for this system, the GBIM rigid body scan provides a reasonable approximation of a free energy surface. We do not expect the GBIM profile to differ significantly had we used GROMOS-96 instead of the OPLS-AA force field to describe the toxin in the rigid body scan.

Results

Toxin Structures in the AT MD Free Energy Simulations. We monitored the structural stability of VSTx1 by calculating the root-mean-squared-deviation (RMSD) of the C_α atoms of the toxin with respect to the initial toxin structure (after a C_α least-squared-fit; Figure S1, Supporting Information). For the majority of the windows, the RMSDs did not exceed 3.5 Å. The only exceptions were toxin locations in water furthest away from the bilayer ($z = -41$ to -39 Å and $z = 39$ to 41 Å) and several toxin locations close to the center of the hydrophobic core of the bilayer ($z = -8$ to -7 Å, $z = 1$ to 3 Å and $z = 4$ and 5 Å) where the RMSDs approached 4 to 5 Å. In the water, VSTx1 experienced a larger conformational drift compared to the ordered lipid environment. For toxin locations close to the center of the hydrophobic core of the bilayer, VSTx1 optimized the interaction of its polar residues with the lipid headgroups resulting in an increase in RMSDs (discussed further below). In all windows, the toxin remained globular and did not unfold. Ignoring the N-terminal residue and the two C-terminal residues of VSTx1 which are flexible,[49] the RMSDs did not exceed 3.5 Å in *any* window (Figure S2, Supporting Information).

We investigated the time-averaged root-mean-squared-fluctuations (RMSF) of the C_α atoms of the toxin with respect to the initial toxin structure (Figures S3 and S4, Supporting Information). It can be seen that the residues near the N- and C-termini (i.e., E1, P33, and F34) had RMSFs approaching and exceeding 3 Å. Consistent with previous AT simulations,[49] four distinct regions with low RMSFs of 0.3 to 1.0 Å were observed between residues 2 and 9, 14 and 16, 19 and 22, and 27 and 30 due to the presence of 3 internal disulfide bridges (between C2 and C16, C9 and C21, and C15 and C28) and 2 β-strands. Consistent with the RMSDs, VSTx1 exhibited greater conformational flexibility in water compared to when buried in the bilayer. The overall pattern of flexibility is consistent with the presence of secondary structure elements and disulfide bridges in the toxin.

Comparison of CG and AT Umbrella Sampling Simulations. We compared the orientation of VSTx1 θ of each AT window vs the corresponding CG window.[20] Figure 3A−C show the distribution of the angle of the hydrophobic moment of the toxin (relative to the bilayer normal; θ) of the AT windows (over 17 to 20 ns; discarding the initial 17 ns as equilibration time) compared with the distributions obtained from the corresponding CG windows (over the final 10 ns, i.e., 30 to 40 ns; after the orientation of VSTx1 had equilibrated).[20] We show this for three windows: (i) with VSTx1 located in water ($z = -41$ to -40 Å; where the com of VSTx1 was restrained at -40.5 Å relative to the com of the bilayer), (ii) with VSTx1 located at the free energy well at the membrane/water interface ($z = -17$ to -16 Å; discussed below), and (iii) with VSTx1 buried close to the center of the hydrophobic core of the bilayer ($z = -2$ to -1 Å).

In water, θ distributions were broad in both CG and AT windows compared to toxin locations in the bilayer, between 40° and 155°. Thus, VSTx1 exhibited greater orientational freedom in water as it tumbled randomly under the influence of the solvent molecules. At the membrane/water interface ($z = -17$ to -16 Å), θ fluctuated over a smaller range of 30° (i.e., between 150° and 180°) in both CG and AT windows. A good overlap can be seen between CG and AT distributions, and θ had an average (± 1 SD) of 168° ± 6° (CG; averaged over 30 to 40 ns) and 168° ± 3° (AT; averaged over 17 to 20 ns), respectively. In this orientation, the polar and hydrophobic residues of VSTx1 were optimally located to interact with the lipid headgroup and tails respectively.

With the toxin buried in the hydrophobic core of the bilayer ($z = -2$ to -1 Å), the CG and AT θ distributions did not overlap; θ had an average of 142° ± 8° and 89° ± 7° in the CG and AT windows, respectively. Inspection of CG window $z = -2$ to -1 Å showed VSTx1 adopted an orientation such that the polar residues of the toxin were positioned to interact with the lipid headgroups of the EC leaflet. Thus, at 0 ns, VSTx1 had a similar orientation in the corresponding AT window (Figure 3C; see inset). Over 20

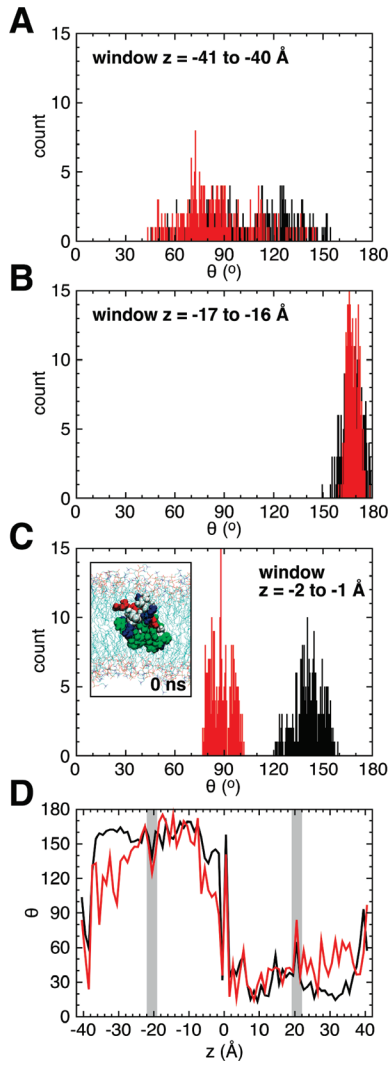


Figure 3. Toxin orientation in the CG and AT simulations. The distributions of the angle of the hydrophobic moment of VSTx1 (relative to the bilayer normal; θ) are shown for CG[20] (black) and AT (red) umbrella sampling windows: (A) $z = -41$ to -40 Å, (B) $z = -17$ to -16 Å, and (C) $z = -2$ to -1 Å. The inset figure in C shows the AT system at 0 ns. (D) Average of θ ($\bar{\theta}$) as a function z for the CG and AT simulations. Averages and distributions are over 30 to 40 ns and 17 to 20 ns per CG and AT window, respectively. The gray regions in D indicate the approximate location of the lipid phosphates.

ns of AT MD, VSTx1 reorientated such that its polar residues were positioned to interact with the lipid headgroups of *both* EC and IC leaflets, accompanied by penetrating water molecules "shielding" the polar residues from the hydrophobic lipid tails. Thus, the hydrophobic surface of VSTx1 was directed perpendicular to the bilayer normal (Figure 1B; right panel). This toxin orientation is possible as VSTx1 was approximately equidistant from the lipid headgroups of both leaflets. Encouragingly, this suggests that 20 ns is sufficient to allow for optimization of the orientation of the toxin in the lipid bilayer in the AT simulations. Thus, the AT simulations showed another mode of interaction of VSTx1 with a POPC bilayer which was not observed in the CG simulations.

In Figure 3D, we plot the average of θ ($\bar{\theta}$) as a function of z in the CG and AT windows. Averages were calculated over 30 to 40 ns and 17 to 20 ns per CG and AT window, respectively. Two distinct values of $\bar{\theta}$ are observed from $z \sim -35$ to 35 Å in the CG simulations: $\sim$155° from $z \sim -35$ to -3 Å and $\sim$25° from $z = 0$ to 35 Å. Thus, VSTx1 underwent a $\sim$180° "flip" as it crossed the bilayer center to minimize the hydrophobic/hydrophilic mismatch between itself and the environment.[20] The "flip-flop" transition in $\bar{\theta}$ between $z \sim -3$ and 0 Å was because chance dictated whether the polar residues of VSTx1 would interact with the headgroups of the EC or IC leaflet. With the toxin located in water in the CG simulations (i.e., $z \sim -41$ to -35 Å and $z \sim 35$ to 41 Å), the SD of θ approached 37° corresponding to VSTx1 tumbling freely in water, compared to SD values of <14° between $z = -35$ and 35 Å when the toxin interacted with the membrane.

In the AT simulations toward water, a somewhat earlier (at $z \sim \pm 25$ Å) and more gradual drift in $\bar{\theta}$ away from the plateau can be seen because of the finer resolution (i.e., more rugged energy landscape) afforded by the AT simulations. We anticipate directionality offered by H-bonding interactions (between VSTx1 and the membrane) in the AT simulations to fine-tune the orientation of the toxin. Between $z \sim -7$ and $+7$ Å with VSTx1 buried in the bilayer, there are clear deviations in the orientation of the toxin over 20 ns, and the difference in $\bar{\theta}$ between the AT and corresponding CG windows was >50°. Thus, between $z = -7$ and $+7$ Å, VSTx1 reorientated such that its polar residues interacted with the lipid headgroups of both EC and IC leaflets. Interaction of the polar residues of Hanatoxin (HATx), a related gating-modifier toxin with an amphipathic molecular surface, with the lipid headgroups of both leaflets, has been reported in recent AT simulations of HATx interacting with a DPPC bilayer.[80] Although our free energy profiles (discussed below) suggest such a mode of interaction is unlikely given the interfacial free energy wells, at least when VSTx1 interacts solely with the bilayer, this may not be physiologically irrelevant when VSTx1 (and other gating-modifier toxins) interacts with both the membrane and the VS of Kv channels.[41,42,47] Overall, $\bar{\theta}$ in the AT simulations correlates with the CG simulations. Taken together, this suggests the CG simulations can be used to guide the setup of the AT simulations, and deviations from the initial AT setup are still permissible (as observed in a few windows with the toxin near the bilayer center) over the time-course of the AT simulations (20 ns).

Free Energy Profiles from Umbrella Sampling MD Simulations. The AT and CG[20] PMF profiles in Figure 4 depict the free energy cost of positioning VSTx1 at different depths in a POPC bilayer. Block analyses (the distributions are split into nonoverlapping blocks to derive multiple PMF profiles in order to evaluate convergence) suggest the AT profile had sufficiently converged (Figure S5, Supporting Information). The AT profile had free energy wells at the membrane/water interface at $\sim\pm$13 to 18 Å, which was somewhat closer to the bilayer center than the CG profile ($\sim\pm$22 Å).[20] The shift in the location of the interfacial minima by $\sim$5 Å between the CG and AT profiles can be accounted for by differences in the equilibrium thickness of the CG and AT membranes and is not due to differences in

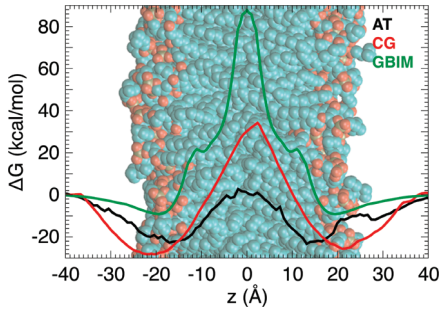


Figure 4. PMF profiles for positioning VSTx1 at different depths in a POPC bilayer. A snapshot of a POPC bilayer is shown in the background for reference. The PMFs shown are derived from AT umbrella sampling simulations (black line), from CG umbrella sampling simulations (red line),[20] and from a rigid body scan of VSTx1 with an implicit solvent and membrane model (GBIM; green line). Because membrane and solvent are implicit and VSTx1 is assumed to be rigid, the GBIM rigid body scan provides a reasonable approximation of a free energy profile. The bilayer center is at $z \sim 0$ Å.

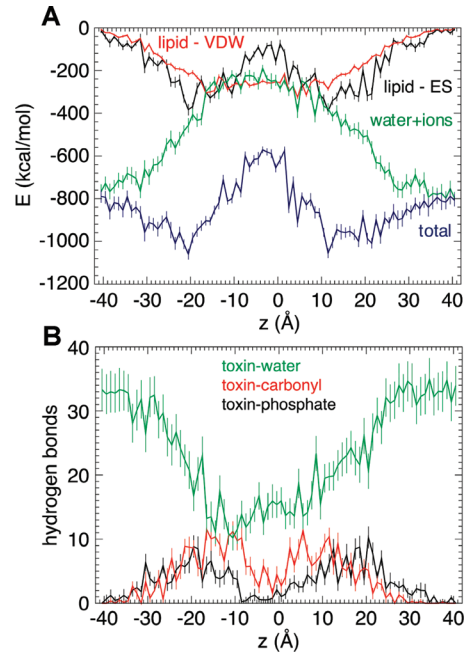


Figure 5. VSTx1/environment interactions in the AT simulations. (A) Toxin−environment interaction (i.e., potential) energies (IE). Average toxin−lipid, toxin−solvent, and total toxin−environment (i.e., toxin−lipid + toxin−solvent) IE. The toxin−lipid IE was decomposed into electrostatic (ES) and VDW components. (B) Hydrogen bonding (H-bonding) interactions. Average number of H bonds between VSTx1 and waters, POPC carbonyls, and POPC phosphates. Averages were taken over 17 to 20 ns/window. Bars represent ±1 SD.

the way VSTx1 interacted with the bilayer. The AT well was −23 kcal/mol (with respect to water), while the central barrier at ∼0 Å was +26 kcal/mol (with respect to the interfacial well). We note the AT profile was not symmetrical about the bilayer center (i.e., $z = 0$ Å). When modeling VSTx1 in the bilayer, we had to remove a few lipids to accommodate the toxin, and this had resulted in asymmetric bilayers (i.e., a bilayer with different numbers of lipids in each leaflet; Figure S6, Supporting Information) and could have contributed to the asymmetry in the free energy profile. Overall, the AT profile suggests it is energetically favorable for VSTx1 to partition into the headgroup/tail interface of a POPC bilayer than to remain in water, and we do not expect the toxin to be able to cross the width of the membrane, consistent with experimental data.[41,42]

Although the topology of the AT and CG profiles was conserved, there were differences in the magnitude of the free energies. The AT and CG free energy wells at the interface were somewhat comparable and differed by ∼4 kcal/mol (−23 vs −27 kcal/mol for AT and CG, respectively). However, the central barrier differed by 36 kcal/mol (+26 vs +61 kcal/mol for AT and CG, respectively, with respect to the well). With VSTx1 buried in the bilayer, water molecules penetrated into the hydrophobic membrane core to provide a micropolar environment for the polar residues of the toxin (Figure 1B, right panel). Water penetration was observed to a more limited extent in the CG simulations.[20] Furthermore, as the polar residues of VSTx1 interacted with the lipid headgroups of both EC and IC leaflets with the toxin buried in the bilayer, this would further stabilize the toxin at the center of the membrane. Taken together, these could account for the increased CG central barrier.

An Approximation of a Free Energy Profile from the Generalized Born Implicit Membrane Model. Although approximate, an implicit membrane and solvent approach allows one to quickly ascertain the topology of the energy landscape and provides a further test of the predicted AT and CG MD PMF profiles. The GBIM rigid body scan provides a reasonable approximation of a free energy profile because membrane and solvent are implicit and VSTx1 is assumed to be rigid. In Figure 4, a free energy well of 9 kcal/mol (with respect to water) was present at ±20 Å, and we observed a central barrier of 97 kcal/mol (with respect to the interfacial well). Thus, the GBIM free energy well was reduced compared to the AT and CG profiles (i.e., −9 vs −23 vs −27 kcal/mol for GBIM, AT, and CG, respectively). The membrane/water interface provides a unique and complex hydrogen bonding (H-bonding) environment that is not sufficiently modeled by an implicit membrane/solvent approach. Enthalpic contributions from H-bonds (i.e., relatively strong electrostatic interactions) between VSTx1, and the lipid phosphate and glycerol moieties (ref 49, and discussed below) at the interface could be significant. Possible water defects in the membrane are modeled to a different extent in the three approaches (i.e., AT, significant; CG, limited; and GBIM, absent), which may explain the differences in the magnitude of the central barrier (i.e., 97 vs 26 vs 61 kcal/mol for GBIM, AT, and CG, respectively). We do not expect the GBIM-derived PMF profile to be very different had we accounted for the internal dynamics of VSTx1 as the toxin has a stable fold.

Toxin-Environment Interaction Energies and Hydrogen Bonding Interactions. To gain a better insight into the nature of the interactions that stabilize VSTx1 in a lipid bilayer environment, we investigated the average interaction (i.e., potential) energies (IE) between the toxin and its environment in the AT simulations. In Figure 5A, we plot

the average toxin−lipid and toxin−solvent IE (averaged over 17 to 20 ns/window) as a function of z, together with the average total toxin−environment IE (i.e., toxin−lipid IE + toxin−solvent IE). The toxin−lipid IE was decomposed into electrostatic (ES) and van der Waals (VDW) components. Only the real space component of PME is reported for the toxin−lipid ES component. The VDW component of the toxin−lipid IE increased (i.e., became more negative) with increased VSTx1 exposure to the membrane towards 0 Å. The ES component of the toxin−lipid IE dictated the topology of the total toxin−environment IE profile. This is consistent with previous AT simulations of VSTx1 with lipid bilayers where ES interactions between the basic residues of the toxin (K4, K8, K10, K17, R24, and K26) and the negatively charged lipid phosphates were important for stabilizing the toxin in the membrane.[49] The toxin−solvent IE is not zero with VSTx1 located at the bilayer center (i.e., at $z = 0$ Å). As discussed earlier, a combination of toxin reorientation, toxin structural drift (as shown by the RMSDs), and penetration of waters ensured continued interaction between VSTx1 and solvent even when the toxin was completely buried in the membrane (Figure 1B, right panel).

In Figure 5B, we investigate H-bonding interactions between VSTx1 and its environment (i.e., waters, lipid carbonyls, and lipid phosphates) in the AT simulations. We plot the average number of H bonds over 17 to 20 ns/window. The toxin−phosphate H bonds had a maximum of 11 at $z \sim \pm 20$ Å, and the toxin−carbonyl H bonds had a maximum of 12 at $z \sim \pm 15$ Å, consistent with the location of the different lipid moieties along the bilayer normal. VSTx1 formed up to 17 hydrogen bonds with waters between $z = -5$ and 5 Å (i.e., when the toxin was located close to the bilayer center). In going from water to the membrane, the loss of H bonds with waters (from, e.g., ∼34 H bonds at ±40 Å) is compensated for by the formation of H bonds with lipids. Indeed, the maximum *total* (i.e., with waters and lipids) number of H bonds is formed when VSTx1 is located at the membrane/water interface at ∼20 Å where, additionally, the hydrophobic residues of the toxin can find a favorable environment as they were directed at the lipid tails. This is consistent with the location of the free energy wells in the PMF profiles. Thus, the free energy of desolvation of the hydrophobic residues of VSTx1 is likely to be a major contributor to the interfacial free energy wells.

Discussion and Conclusion

We have combined CG and AT simulations serially to compute a 1D PMF (i.e., free energy) profile for the interaction of a small protein with a lipid bilayer. We used information gained from CG free energy simulations[20] to guide the setup of corresponding AT simulations. We calculated the 1D free energy profile of VSTx1 interacting with a POPC bilayer, where the reaction coordinate corresponds to the position, projected along the bilayer normal, of the com of the toxin with respect to the com of the membrane. The VSTx1/bilayer system was chosen as a test system for multiscale analysis of protein/membrane interactions as it is reasonably simple and well characterized.[20,41,42,49] How VSTx1 interacts with membranes is also of interest in the context of the biophysics of voltage sensing by potassium channels.[41,42,47,81−83] The interaction of VSTx1 with a bilayer will also provide insights into how small amphipathic proteins and peptides interact with lipid bilayers. From a theoretical point of view, in a 1D PMF of VSTx1 along the membrane normal z, there is an important nonsampled degree of freedom that had to equilibrate in order to yield meaningful PMFs. The important nonsampled degree of freedom is due to the amphipathic surface of VSTx1, giving the toxin a distinct orientation as a function of position in the bilayer. The longer time scales accessible to CG simulations allowed the toxin orientation to equilibrate. The initial AT configurations were based on the equilibrated CG configurations, and the AT simulations provided a finer view of how VSTx1 interacted with the membrane.

The CG and AT PMF profiles had a conserved topology (i.e., interfacial wells and a central barrier), but there were differences in the magnitude of the free energies. The interfacial free energy wells are comparable (i.e., −27 vs −23 kcal/mol for CG and AT, respectively), which remains considerably larger than that derived experimentally for VSTx1 (−7 kcal/mol[48]) and related gating-modifier toxins that target the VS of Kv channels (−3.5 to −8.5 kcal/mol[84]). With VSTx1 located close to the interfacial well, e.g., in AT window $z = -17$ to -16 Å, the com of L30 of the toxin, centered on the hydrophobic patch, was positioned at a distance of ∼10 Å from the bilayer center, which is in agreement with depth-dependent fluorescence quenching data on HATx (W30 of HATx1 was reported to be positioned at a distance of 9 Å from the membrane center).[47] The largest difference between the CG and AT profiles was the central barrier (61 vs 26 kcal/mol for CG and AT, respectively). This difference may be accounted for by phenomena observed in the AT simulations which were absent or observed to a limited extent in the CG simulations: (1) water defects in the membrane (which were more limited in the CG simulations because of the lack of dipoles in CG waters[7]) and (2) when displaced toward the membrane center, VSTx1 reorientated such that its polar residues interacted with the lipid headgroups of *both* leaflets, consistent with simulations of a related toxin[80] (this was not observed in the CG simulations[20]). We emphasize that this 1D PMF profile would be difficult to compute using AT simulations in isolation, without guidance from CG simulations, because of likely problems with sampling and convergence.

To illustrate this, we performed additional simulations of VSTx1 located at the interfacial free energy well but in a nonoptimal initial orientation (i.e., "upside-down" relative to the bilayer; Figure S9, Supporting Information). The toxin would require well in excess of 20 ns to locate its optimal orientation in the bilayer. It would therefore be difficult to compute this PMF, using an AT force-field, without *a priori* knowledge of toxin orientation as a function of reaction coordinate z.

Despite the CG bias, for a few toxin locations close to the bilayer center that we observed, VSTx1 was able to deviate substantially from its initial orientation in the AT simulations to adopt configurations which were not observed in the CG simulations. Thus, the AT simulations could

sample new configurations of toxin orientation. Taken together, it is clear a serial multiscale approach has allowed for better estimates of this PMF profile. One could also combine the multiscale procedure presented here with other enhanced sampling methods (e.g., hybrid Monte Carlo[85] or replica exchange[86]) to achieve further exploration of the AT energy landscape.

The AT simulations provided insights into how lipids might interact with gating-modifier toxins such as VSTx1 when they are located at their optimal binding depth in lipid bilayers. The lipids were seen to "wrap" their acyl chains around the hydrophobic face of VSTx1 (Figure S8, Supporting Information).

Returning to a more biological perspective, it is of interest that recent experimental studies of the action of VSTx1 on Kv channels have been interpreted in terms of perturbation of membrane/channel forces by the toxin.[82] Thus, a detailed understanding of the nature and location of the toxin/bilayer interaction becomes crucial to our understanding of the mode of action of the toxin. The multiscale procedure presented here provides a valuable tool for such studies.

Acknowledgment. We thank our colleagues in the Structural Bioinformatics and Computational Biochemistry Unit, in particular Ranjit Vijayan, for their helpful comments. C.L.W. is funded by the Oxford Centre for Integrative Systems Biology (OCISB). Research in M.S.P.S.'s laboratory is funded by BBSRC and the Wellcome Trust. We acknowledge the U.K. National Grid Service for computing resources.

Supporting Information Available: Additional analyses of the simulations described in this paper. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Rodinger, T.; Pomès, R. Enhancing the accuracy, the efficiency and the scope of free energy simulations. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164–170.

(2) Allen, T. W.; Andersen, O. S.; Roux, B. Energetics of ion conduction through the gramicidin channel. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117–122.

(3) Bernèche, S.; Roux, B. Energetics of ion conduction through the K^+ channel. *Nature* **2001**, *414*, 73–77.

(4) Hub, J. S.; de Groot, B. L. Mechanism of selectvity in aquaporins and aquaglyceroporins. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1198–1203.

(5) Pongprayoon, P.; Beckstein, O.; Wee, C. L.; Sansom, M. S. P. Simulations of anion transport through OprP reveal the molecular basis for high affinity and selectivity for phosphate. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 21614–21618.

(6) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.* **2008**, *94*, 3393–3404.

(7) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(8) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matt.* **2004**, *16*, R481–R512.

(9) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(10) Shi, Q.; Izvekov, S.; Voth, G. A. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane bound ion channel. *J. Phys. Chem. B* **2006**, *110*, 15045–15048.

(11) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. Coarse grained protein-lipid model with application to lipoprotein particles. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.

(12) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.* **2007**, *157*, 593–605.

(13) Marrink, S. J.; Risselada, J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI forcefield: coarse grained model for biomolecular simulations. *J. Phys. Chem. B.* **2007**, *111*, 7812–7824.

(14) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: Boca Raton, FL, 2008.

(15) Periole, X.; Cavalli, M.; Marrink, S. J.; Ceruso, M. A. Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *J. Chem. Theory Comput.* **2009**, *5*, 2531–2543.

(16) Scott, K. A.; Bond, P. J.; Ivetac, A.; Chetwynd, A. P.; Khalid, S.; Sansom, M. S. P. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure* **2008**, *16*, 621–630.

(17) Sansom, M. S. P.; Scott, K. A.; Bond, P. J. Coarse grained simulation: a high throughput computational approach to membrane proteins. *Biochem. Soc. Trans.* **2008**, *36*, 27–32.

(18) Wee, C. L.; Bemporad, D.; Sands, Z. A.; Gavaghan, D.; Sansom, M. S. P. SGTx1, a Kv channel gating-modifier toxin, binds to the interfacial region of lipid bilayers. *Biophys. J.* **2007**, *92*, L07–L09.

(19) Wee, C. L.; Balali-Mood, K.; Gavaghan, D.; Sansom, M. S. P. The interaction of phospholipase A2 with a phospholipid bilayer: coarse-grained molecular dynamics simulations. *Biophys. J.* **2008**, *95*, 1649–1657.

(20) Wee, C. L.; Gavaghan, D.; Sansom, M. S. P. Lipid bilayer deformation and the free energy of interaction of a K_v channel gating-modifier toxin. *Biophys. J.* **2008**, *95*, 3816–3826.

(21) Periole, X.; Huber, T.; Marrink, S. J.; Sakmar, T. P. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.

(22) Bond, P. J.; Sansom, M. S. P. Bilayer deformation by the Kv channel voltage sensor domain revealed by self-assembly simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2631–2636.

(23) Venturoli, M.; Smit, B.; Sperotto, M. M. Simulation studies of protein-induced bilayer deformations, and lipid-induced protein tilting, on a mesoscopic model for lipid bilayers with embedded proteins. *Biophys. J.* **2005**, *88*, 1778–1798.

(24) Treptow, W.; Marrink, S.-J.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.

(25) Yefimov, S.; van der Giessen, E.; Onck, P. R.; Marrink, S. J. Mechanosensitive membrane channels in action. *Biophys. J.* **2008**, *94*, 2994–3002.

(26) Hall, B. A.; Sansom, M. S. P. Coarse-grained MD simulations and protein-protein interactions: the cohesin-dockerin system. *J. Chem. Theory Comput.* **2009**, *5*, 2465–2471.

(27) Lindahl, E.; Sansom, M. S. P. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 425–431.

(28) Bond, P. J.; Wee, C. L.; Sansom, M. S. P. Coarse-grained molecular dynamics simulations of the energetics of helix insertion into a lipid bilayer. *Biochemistry* **2008**, *47*, 11321–11331.

(29) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse grained force field: extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(30) Psachoulia, E.; Bond, P. J.; Fowler, P. W.; Sansom, M. S. P. Helix-helix interactions in membrane proteins: coarse grained simulations of glycophorin helix dimerization. *Biochem.* **2008**, *47*, 10503–105012.

(31) Allen, T. W. Modeling charged protein side chains in lipid membranes. *J. Gen. Physiol.* **2007**, *130*, 237–240.

(32) Vorobyov, I.; Li, L.; Allen, T. W. Assessing atomistic and coarse-grained force fields for protein-lipid interactions: the formidable challenge of an ionizable side chain in a membrane. *J. Phys. Chem. B* **2008**, *112*, 9588–9602.

(33) Chang, R.; Ayton, G. S.; Voth, G. A. Multiscale coupling of mesoscopic- and atomistic-level lipid bilayer simulations. *J. Chem. Phys.* **2005**, *122*, 244716.

(34) Ayton, G. A.; Noid, W. G.; Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(35) Ayton, G. S.; Voth, G. A. Multiscale simulation of transmembrane proteins. *J. Struct. Biol.* **2007**, *157*, 570–578.

(36) Ayton, G. S.; Voth, G. A. Systematic multiscale simulation of membranes protein systems. *Curr. Opin. Struct. Biol.* **2009**, *19*, 138–144.

(37) Carpenter, T.; Bond, P. J.; Khalid, S.; Sansom, M. S. P. Self-assembly of a simple membrane protein: coarse-grained molecular dynamics simulations of the influenza M2 channel. *Biophys. J.* **2008**, *95*, 3790–3801.

(38) Orsi, M.; Sanderson, W. E.; Essex, J. W. Permeability of small molecules through a lipid bilayer: a multiscale simulation study. *J. Phys. Chem. B* **2009**, *113*, 12019–12029.

(39) Ruta, V.; Jiang, Y. X.; Lee, A.; Chen, J. Y.; MacKinnon, R. Functional analysis of an archaebacterial voltage-dependent K^+ channel. *Nature* **2003**, *422*, 180–185.

(40) Jiang, Y.; Ruta, V.; Chen, J.; Lee, A. G.; Mackinnon, R. The principle of gating charge movement in a voltage-dependent K^+ channel. *Nature* **2003**, *423*, 42–48.

(41) Swartz, K. J. Tarantula toxins interacting with voltage sensors in potassium channels. *Toxicon* **2007**, *49*, 213–230.

(42) Milescu, M.; Vobecky, J.; Roh, S. H.; Kim, S. H.; Jung, H. J.; Il Kim, J.; Swartz, K. J. Tarantula toxins interact with voltage sensors within lipid membranes. *J. Gen. Physiol.* **2007**, *130*, 497–511.

(43) Takahashi, H.; Kim, J. I.; Min, H. J.; Sato, K.; Swartz, K. J.; Shimada, I. Solution structure of hanatoxin1, a gating modifier of voltage- dependent K^+ channels: Common surface features of gating modifier toxins. *J. Mol. Biol.* **2000**, *297*, 771–780.

(44) Jung, H. J.; Lee, J. Y.; Kim, S. H.; Eu, Y. J.; Shin, S. Y.; Milescu, M.; Swartz, K. J.; Kim, J. I. Solution structure and lipid membrane partitioning of VSTx1, an inhibitor of the KvAP potassium channel. *Biochemistry* **2005**, *44*, 6015–6023.

(45) Lee, C. W.; Kim, S.; Roh, S. H.; Endoh, H.; Kodera, Y.; Maeda, T.; Kohno, T.; Wang, J. M.; Swartz, K. J.; Kim, J. I. Solution structure and functional characterisation of SGTx1, a modifier of Kv2.1 channel gating. *Biochemistry* **2004**, *43*, 890–897.

(46) Jung, H. J.; Lee, J. Y.; Kim, S. H.; Eu, Y. J.; Shin, S. Y.; Milescu, M.; Swartz, K. J.; Kim, J. I. Solution structure and lipid membrane partitioning of VSTx1, an inhibitor of the KvAP potassium channel. *Biochemistry* **2005**, *44*, 6015–6023.

(47) Phillips, L. R.; Milescu, M.; Li-Smerin, Y.; Midell, J. A.; Kim, J. I.; Swartz, K. J. Voltage-sensor activation with a tarantula-toxin as cargo. *Nature* **2005**, *436*, 857–860.

(48) Lee, S. Y.; MacKinnon, R. A membrane-access mechanism of ion channel inhibition by voltage sensor toxins from spider venom. *Nature* **2004**, *430*, 232–235.

(49) Bemporad, D.; Wee, C. L.; Sands, Z.; Grottesi, A.; Sansom, M. S. P. VSTx1, a modifier of Kv channel gating, localizes to the interfacial region of lipid bilayers. *Biochemistry* **2006**, *45*, 11844–11855.

(50) Ulmschneider, M. B.; Sansom, M. S. P.; Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 252–265.

(51) Dominy, B. N.; Brooks, C. L. Development of a generalized Born model parameterisation for proteins and nucleic acids. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(52) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(53) Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

(54) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte-Carlo free energy distributions: umbrella sampling. *J Comput. Phys.* **1977**, *23*, 187–199.

(55) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(56) Wang, J. M.; Roh, S. H.; Kim, S.; Lee, C. W.; Kim, J. I.; Swartz, K. J. Molecular surface of tarantula toxins interacting with voltage sensors in Kv channels. *J. Gen. Physiol.* **2004**, *123*, 455–467.

(57) Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **1984**, *179*, 125–142.

(58) Krznaric, D.; Levcopoulos, C. Fast algorithms for complete linkage clustering. *Discrete Comp. Geom.* **1998**, *19*, 131–145.

(59) Bond, P. J.; Sansom, M. S. P. Insertion and assembly of membrane proteins via simulation. *J. Am. Chem. Soc.* **2006**, *128*, 2697–2704.

(60) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.

(61) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(62) van Gunsteren, W. F.; Kruger, P.; Billeter, S. R.; Mark, A. E.; Eising, A. A.; Scott, W. R. P.; Huneberger, P. H.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Biomos & Hochschulverlag AG an der ETH Zurich: Groningen; Zurich, 1996.

(63) Berger, O.; Edholm, O.; Jahnig, F. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidycholine at full hydration, constant pressure and constant temperature. *Biophys. J.* **1997**, *72*, 2002–2013.

(64) Marrink, S. J.; Berger, O.; Tieleman, D. P.; Jahnig, F. Adhesion forces of lipids in a phospholipid membrane studied by molecular dynamics simulations. *Biophys. J.* **1998**, *74*, 931–943.

(65) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. A consistent empirical potential for water-protein interactions. *Biopolymers* **1984**, *23*, 1513–1518.

(66) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(67) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(68) Miyamoto, S.; Kollman, P. A. Settle - an analytical version of the Shake and Rattle algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(69) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(70) Ulmschneider, J. P.; Ulmschneider, M. B. Folding simulations of the transmembrane helix of Virus Protein U in an implicit membrane model. *J. Chem. Theory Comput.* **2007**, *3*, 2335–2346.

(71) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, C. W. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(72) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. Free energies of hydration from a generalized Born model and an ALL-atom force field. *J. Phys. Chem. B* **2004**, *108*, 16264–16270.

(73) Parsegian, A. Energy of an ion crossing a low dielectric membrane: Solution to four relevant electrostatics problems. *Nature* **1969**, *221*, 844–846.

(74) Spassov, V. Z.; Yan, L.; Szalma, S. Introducing an implicit membrane in generalized Born/solvent accessibility continuum solvent models. *J. Phys. Chem. B* **2002**, *106*, 8726–2738.

(75) Killian, J. A. Synthetic peptides as models for intrinsic membrane proteins. *FEBS Lett.* **2003**, *555*, 134–138.

(76) de Planque, M. R. R.; Killian, J. A. Protein-lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring. *Mol. Membr. Biol.* **2003**, *20*, 271–284.

(77) Wiener, M. C.; White, S. H. Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. III. Complete structure. *Biophys. J.* **1992**, *61*, 434–447.

(78) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(79) Radzicka, A.; Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **1988**, *27*, 1664–1670.

(80) Nishizawa, M.; Nishizawa, K. Interaction between K^+ channel gate modifier hanatoxin and lipid bilayer membranes analyzed by molecular dynamics simulation. *Eur. Biophys. J.* **2006**, *35*, 373–381.

(81) Alabi, A. A.; Bahamonde, M. I.; Jung, H. J.; Kim, H. J.; Swartz, K. J. Portability of paddle motif function and pharmacology in voltage sensors. *Nature* **2007**, *450*, 370–375.

(82) Schmidt, D.; MacKinnon, R. Voltage-dependent K^+ channel gating and voltage sensor toxin sensitivity depend on the mechanical state of the lipid membrane. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 19276–19281.

(83) Swartz, K. J. Sensing voltage across lipid membranes. *Nature* **2008**, *456*, 891–897.

(84) Posokhov, Y. O.; Gottlieb, P. A.; Morales, M. J.; Sachs, F.; Ladokhin, A. S. Is lipid bilayer binding a common property of inhibitor cysteine knot ion-channel blockers? *Biophys. J.* **2007**, *93*, L20–L22.

(85) Wee, C. L.; Sansom, M. S. P.; Reich, S.; Akhmatskaya, E. Improved sampling for simulations of interfacial membrane proteins: application of Generalized Shadowing Hybrid Monte Carlo to a peptide toxin/bilayer system. *J. Phys. Chem. B* **2008**, *112*, 5710–5717.

(86) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

CT900652S

JCTC Journal of Chemical Theory and Computation

Erratum

Rigorous Extraction of the Anisotropic Multispin Hamiltonian in Bimetallic Complexes from the Exact Electronic Hamiltonian. [*J. Chem. Theory Comput. 6,* 55–65 (2010)]. By Rémi Maurice,* Nathalie Guihéry, Roland Bastardis, and Coen de Graaf.

Table 4. Two typographical errors are worth noting:

• The matrix elements $\langle 2, -2|H_{mod}|0, 0\rangle$, $\langle 2, 2|H_{mod}|0, 0\rangle$, $\langle 0, 0|H_{mod}|2, -2\rangle$, and $\langle 0, 0|H_{mod}|2, 2\rangle$, that were reported as $[2/(\sqrt{3})](E_a - E_{ab})$ in the published version, are equal to $[1/(\sqrt{3})](2E_a - E_{ab})$.

• The matrix elements $\langle 1, -1|H_{mod}|1, 1\rangle$ and $\langle 1, 1|H_{mod}|1, -1\rangle$, that were reported as $-E_a - E_{ab}$ in the published version, are equal to $-E_a + E_{ab}$.

The entire corrected version of the matrix is presented. All results and exploitations reported in the article used the correct expressions.

Table 4. Matrix Elements of the Model Hamiltonian for Bimetallic Ni(II) Complexes with Magnetic Anisotropy in the Coupled $|S, M_S\rangle$ Basis

$\|S, M_S\rangle$	$\|2, -2\rangle$	$\|2, -1\rangle$	$\|2, 0\rangle$	$\|2, 1\rangle$	$\|2, 2\rangle$
$\langle 2, -2\|$	$J + (2/3)(D_a + D_{ab})$	0	$[\sqrt{(2/3)}](E_a + E_{ab})$	0	0
$\langle 2, -1\|$	0	$J - (1/3)(D_a + D_{ab})$	0	$E_a + E_{ab}$	0
$\langle 2, 0\|$	$[\sqrt{(2/3)}](E_a + E_{ab})$	0	$J - (2/3)(D_a + D_{ab})$	0	$[\sqrt{(2/3)}](E_a + E_{ab})$
$\langle 2, 1\|$	0	$E_a + E_{ab}$	0	$J - (1/3)(D_a + D_{ab})$	0
$\langle 2, 2\|$	0	0	$[\sqrt{(2/3)}](E_a + E_{ab})$	0	$J + (2/3)(D_a + D_{ab})$
$\langle 1, -1\|$	0	0	0	0	0
$\langle 1, 0\|$	0	0	0	0	0
$\langle 1, 1\|$	0	0	0	0	0
$\langle 0, 0\|$	$[1/(\sqrt{3})](2E_a - E_{ab})$	0	$[(\sqrt{2})/3](2D_a - D_{ab})$	0	$[1/(\sqrt{3})](2E_a - E_{ab})$
	$\|1, -1\rangle$	$\|1, 0\rangle$	$\|1, 1\rangle$	$\|0, 0\rangle$	
$\langle 2, -2\|$	0	0	0	$[1/(\sqrt{3})](2E_a - E_{ab})$	
$\langle 2, -1\|$	0	0	0	0	
$\langle 2, 0\|$	0	0	0	$[(\sqrt{2})/3](2D_a - D_{ab})$	
$\langle 2, 1\|$	0	0	0	0	
$\langle 2, 2\|$	0	0	0	$[1/(\sqrt{3})](2E_a - E_{ab})$	
$\langle 1, -1\|$	$-J - (1/3)(D_a - D_{ab})$	0	$-E_a + E_{ab}$	0	
$\langle 1, 0\|$	0	$-J + (2/3)(D_a - D_{ab})$	0	0	
$\langle 1, 1\|$	$-E_a + E_{ab}$	0	$-J - (1/3)(D_a - D_{ab})$	0	
$\langle 0, 0\|$	0	0	0	$-2J$	

CT100053G

10.1021/ct100053g
Published on Web 02/16/2010

Published on Web 02/16/2010