JCTC Journal of Chemical Theory and Computation

Application of the Electrostatically Embedded Many-Body Expansion to Microsolvation of Ammonia in Water Clusters

Anastassia Sorkin, Erin E. Dahlke, and Donald G. Truhlar*

Department of Chemistry and Supercomputing Institute,
University of Minnesota,
Minneapolis, Minnesota 55403-0431

Received December 21, 2007

Abstract: The electrostatically embedded many-body expansion (EE-MB), at both the second and third order, that is, the electrostatically embedded pairwise additive (EE-PA) approximation and the electrostatically embedded three-body (EE-3B) approximation, are tested for mixed ammonia−water clusters. We examine tetramers, pentamers, and hexamers for three different density functionals and two levels of wave function theory, We compare the many-body results to the results of full calculations performed without many-body expansions. Because of the differing charge distributions in the two kinds of monomers, this provides a different kind of test of the usefulness of the EE-MB method than was provided by previous tests on pure water clusters. We find only small errors due to the truncation of the many-body expansion for the mixed clusters. In particular, for tests on tetramers and pentamers, the mean absolute deviations for truncation at second order are 0.36−0.98 kcal/mol (average: 0.66 kcal/mol), and the mean absolute deviations for truncation at third order are 0.04−0.28 (average: 0.16 kcal/mol). These may be compared to a spread of energies as large as 4.24 kcal/mol in the relative energies of various structures of pentamers and to deviations of up to 8.57 kcal/mol of the full calculations of relative energies from the best estimates of the relative energies. When the methods are tested on hexamers, the mean unsigned deviation per monomer remains below 0.10 kcal/mol for EE-PA and below 0.03 kcal/mol for EE-3B. Thus the additional error due to the truncation of the expansion is small compared to the accuracy needed or the other approximations involved in practical calculations. This means that the EE-MB expansion in combination with density functional theory or wave function theory for the oligomers provides a useful practical model chemistry for making electronic structure calculations and simulations more affordable by improving the scaling with respect to system size.

1. Introduction

The accurate calculation of energies and other characteristics of large systems is a challenging physical and chemical problem. Molecular mechanics is able to treat large systems but does not provide sufficient accuracy for many problems. On the other hand, more accurate post Hartree−Fock methods (such as MP2, CCSD, or CCSD(T)) and density functional theory (DFT) are very expensive (relative to molecular mechanics) and their computational cost increases quickly with the system size (for example CCSD(T) scales as N^7, where N is the number of atoms in this system), which make applications of these methods to very large systems impractical.

Recently, the electrostatically embedded many-body (EE-MB) method was developed.[1] This method is similar to methods developed by Kitaura and Fedorov[2,3] and Hirata et al.[4] but is easier to apply; in particular the method was formulated in such a way that it is very easy to calculate energy gradients. In the EE-MB method, the system is divided into fragments (typically dimers or trimers), and each fragment is treated in a field of point charges representing the electrostatic potential of the other fragments. The locations of the point charges depend on the geometry of the other fragments, but the magnitudes do not. (Generalizations to more complicated prescriptions for the point charges are possible but will not be considered in the present work.) It was shown in previous work[1,5] that the three-body version (EE-3B) of EE-MB with this simple prescription for the electrostatics, when applied to water clusters containing 5−21 water molecules, yields a mean error less than 0.4 kcal/mol for all levels of theory examined.

Small $NH_3(H_2O)_n$ clusters play an important role in atmospheric aerosol formation.[6,7] $NH_3(H_2O)_n$ and similar complexes have been extensively studied both experimentally[8–11] and theoretically,[12–16] and in the present article we have applied the EE-MB method to calculate total energies of such clusters with $n = 3−5$. The calculations were done with three density functionals (PBE, B3LYP, and M06-2X) and two wave function methods (MP2 and CCSD(T)) with various basis sets. The accuracies of the EE-PA (two-body) and EE-3B (three-body)

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

10.1021/ct7003462 CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/19/2008

Table 1. Sets of Charges Used in the Present Calculations

	ammonia		water	
	q_N	q_H	q_O	q_H
M1	−0.867	0.289	−0.690	0.345
M2	−0.908	0.303	−0.716	0.358
M3	−0.953	0.318	−0.749	0.374
CM4	−0.906	−0.302	−0.648	0.324

Table 2. Relative Energies (kcal/mol) Predicted by Full Calculations of Tetramers: $NH_3(H_2O)_3$[c]

	B	C	D	E
PBE/DZ	5.01	6.27	14.31	9.11
PBE/TZ	3.77	5.28	12.07	7.77
B3LYP/DZ	4.29	5.75	7.42	7.91
B3LYP/6-311+G(d,p)[a]	3.97	5.75	6.91	7.36
B3LYP/TZ	3.30	4.80	6.23	6.78
M06-2X/DZ	3.80	3.72	6.92	7.28
M06-2X/TZ	2.60	2.90	5.64	6.10
MP2/DZ	4.15	4.75	6.83	7.19
MP2/6-311+G(d,p)[a]	4.03	4.70	6.60	7.00
MP2/TZ	3.32	3.91	6.02	6.53
CCSD(T)/DZ	3.87	4.58	6.45	6.85
QCISD(T)/6-311+G(d,p)[a]	3.79	4.52	6.32	6.74
extrapolated[b]	3.08	3.73	5.74	6.27

[a] Reference 16. [b] Best estimate obtained by eq 1. [c] All energies are relative to the isomer A.

Table 3. Relative Energies (kcal/mol) Predicted by Full Calculations of Pentamers: $NH_3(H_2O)_4$[c]

	B	C	D	E
PBE/DZ	2.06	0.95	2.62	4.24
PBE/TZ	1.15	0.65	2.12	3.43
B3LYP/DZ	2.09	0.92	2.40	4.18
B3LYP/6-311+G(d,p)[a]	1.94	0.77	2.25	4.12
B3LYP/TZ	1.12	0.003	2.01	3.36
M06-2X/DZ	1.71	0.79	1.17	3.41
M06-2X/TZ	1.50	0.94	1.85	3.37
MP2/DZ	0.80	0.66	1.52	3.29
MP2/6-311+G(d,p)[a]	0.75	0.85	1.59	3.33
MP2/TZ	−0.09	0.14	0.97	2.34
CCSD(T)/DZ	0.84	0.68	1.52	3.46
QCISD(T)/6-311+G(d,p)[a]	0.42	0.69	1.37	3.18
extrapolated[b]	−0.42	−0.02	0.75	2.19

[a] Reference 13. [b] Best estimate obtained by eq 1. [c] All energies are relative to the isomer A.

Table 4. Mean Unsigned Deviations (kcal/mol) from Full Calculations for Four Models of Point Charges for Five Tetramers $NH_3(H_2O)_3$ and Five Pentamers $NH_3(H_2O)_4$[a]

	M1	M2	M3	CM4
	EE-PA			
PBE/DZ	0.85	0.75	0.63	0.98
PBE/TZ	0.72	0.61	0.50	0.45
B3LYP/DZ	0.81	0.69	0.58	0.93
B3LYP/TZ	0.64	0.54	0.47	0.77
M06-2X/DZ	0.71	0.62	0.52	0.83
M06-2X/TZ	0.54	0.48	0.36	0.71
MP2/DZ	0.76	0.66	0.60	0.85
MP2/TZ	0.71	0.62	0.51	0.83
	EE-3B			
PBE/DZ	0.21	0.20	0.18	0.22
PBE/TZ	0.23	0.22	0.20	0.24
B3LYP/DZ	0.12	0.12	0.11	0.12
B3LYP/TZ	0.18	0.19	0.19	0.17
M06-2X/DZ	0.27	0.26	0.26	0.28
M06-2X/TZ	0.16	0.15	0.14	0.17
MP2/DZ	0.08	0.07	0.06	0.08
MP2/TZ	0.06	0.05	0.04	0.06

[a] Deviations in absolute electronic energies averaged over ten structures.

Table 5. Mean Unsigned Errors (in kcal/mol) from Full Calculations of the Relative Energies for Four Models of Point Charges for Tetramers $NH_3(H_2O)_3$[a]

	M1	M2	M3	CM4
	EE-PA			
PBE/DZ	0.85	0.79	0.72	0.92
PBE/TZ	0.86	0.79	0.73	0.92
B3LYP/DZ	0.42	0.39	0.35	0.47
B3LYP/TZ	0.39	0.35	0.31	0.43
M06-2X/DZ	0.31	0.28	0.24	0.34
M06-2X/TZ	0.24	0.21	0.17	0.28
MP2/DZ	0.25	0.18	0.26	0.29
MP2/TZ	0.25	0.22	0.18	0.28
CCSD(T)/DZ	0.28	0.25	0.24	0.32
	EE-3B			
PBE/DZ	0.13	0.12	0.11	0.13
PBE/TZ	0.12	0.12	0.11	0.13
B3LYP/DZ	0.05	0.04	0.04	0.05
B3LYP/TZ	0.05	0.05	0.04	0.06
M06-2X/DZ	0.09	0.09	0.08	0.09
M06-2X/TZ	0.08	0.07	0.07	0.08
MP2/DZ	0.03	0.03	0.04	0.04
MP2/TZ	0.02	0.02	0.01	0.02
CCSD(T)/DZ	0.04	0.04	0.04	0.05

[a] Averaged over ten pairs of structures.

versions of EE-MB were ascertained by comparison to full calculations, i.e., calculations that do not employ a many-body expansion.

2. Computational Details

All calculations in this paper were carried out using the *MN-GFM* module version 3.0 for incorporation of new DFT models into *Gaussian03*.[17]

The starting geometries of ammonia−water clusters $NH_3(H_2O)_n$ with $n = 3,4$ were taken from Bacelo,[16] who optimized them at the MP2/6-311+G** level. We reoptimized these structures with the PBE,[18] B3LYP,[19–22] and M06-2X[23] functionals and the MP2[24] post Hartree−Fock method with the 6-311++G(2d,2p) basis set. (The M06-2X functional was recently found to be very accurate for main-group thermochemistry and kinetics and for noncovalent interactions.[23]) EE-PA and EE-3B calculations were carried out with each of the methods with geometries optimized by the full calculations with the corresponding method. The calculations, including new geometry optimizations, were then repeated with the 6-31+G(d,p) basis set. EE-PA and EE-3B calculations with the CCSD(T)/6-31+G(d,p)[25] calculations were performed with the geometries optimized with B3LYP/6-311++G(2d, 2p) method.

Bacelo[16] noted that the geometry of $NH_3(H_2O)_n$ clusters with small n are similar to the geometries of stable water clusters. Therefore two starting geometries of $NH_3(H_2O)_5$ were generated from the geometries of water hexamers (cage and ring), taken from the *The Cambridge Cluster Database*,[26] and one water molecule in each hexamer was replaced by an ammonia molecule. The resulting structures were optimized with the PBE, B3LYP, and M06-2X functionals and the MP2 method, all with

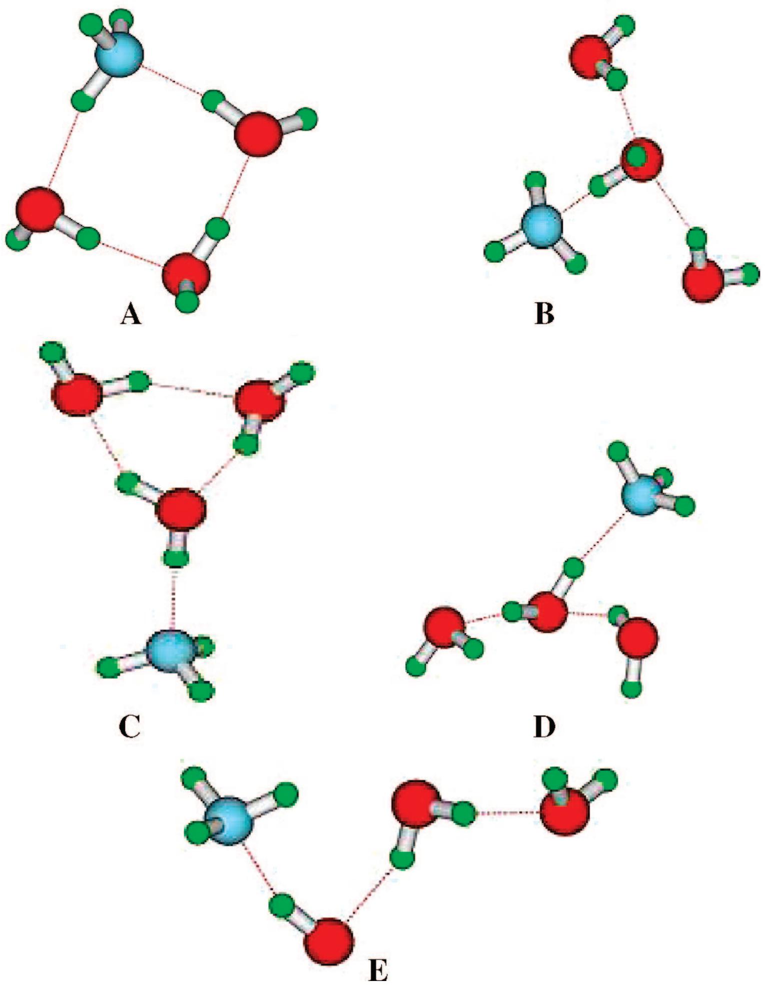


Figure 1. Structural isomers of $NH_3(H_2O)_3$ (tetramers) used in this work.

Table 6. Mean Unsigned Errors (kcal/mol) from Full Calculations of the Relative Energies for Four Models of Point Charges in Pentamers $NH_3(H_2O)_4$[a]

	M1	M2	M3	CM4
		EE-PA		
PBE/DZ	0.69	0.67	0.64	0.68
PBE/TZ	0.79	0.78	0.76	0.81
B3LYP/DZ	0.63	0.64	0.65	0.62
B3LYP/TZ	0.41	0.36	0.60	0.44
M06-2X/DZ	0.30	0.30	0.30	0.28
M06-2X/TZ	0.28	0.28	0.27	0.29
MP2/DZ	0.30	0.29	0.27	0.32
MP2/TZ	0.18	0.17	0.16	0.21
		EE-3B		
PBE/DZ	0.21	0.20	0.18	0.22
PBE/TZ	0.23	0.21	0.22	0.23
B3LYP/DZ	0.31	0.32	0.31	0.32
B3LYP/TZ	0.16	0.16	0.16	0.16
M06-2X/DZ	0.11	0.10	0.10	0.11
M06-2X/TZ	0.10	0.10	0.09	0.11
MP2/DZ	0.09	0.08	0.08	0.08
MP2/TZ	0.05	0.04	0.04	0.05

[a] Averaged over ten pairs of structures.

Table 7. Mean Unsigned Deviations (kcal/mol) for Hexamers $NH_3(H_2O)_5$ for Four Models of Point Charges in the EE-3B Approximation Calculated with the TZ Basis Set[a]

	M1	M2	M3	CM4
		EE-PA		
PBE	0.66	0.58	0.53	0.70
B3LYP	0.75	0.73	0.61	0.86
M06-2X	0.55	0.50	0.39	0.63
MP2	0.62	0.59	0.49	0.69
		EE-3B		
PBE	0.35	0.34	0.33	0.37
B3LYP	0.09	0.09	0.06	0.10
M06-2X	0.35	0.34	0.34	0.36
MP2	0.09	0.09	0.08	0.10

[a] Deviations in absolute electronic energies averaged over two structures.

the 6-311++G(2d,2p) basis set. EE-PA and EE-3B calculations were also carried out for these hexamers.

The EE-PA and EE-3B calculations were tested with 4 sets of charges listed in Table 1. The three sets of charges M1, M2, and M3 were calculated respectively with PBE, B3LYP, and MP2, in each case using the ChelpG scheme[27] on the monomers and the 6-311++G(2d,2p) basis set. The last set of charges, CM4, was calculated with the CM4 charge model[28] using B3LYP/6-31+G(d,p) on monomers. In our original test[2] (for pure water clusters), we found that the best results were obtained with $q_O = -0.778$ and $q_O = -0.834$. Of the four sets of charges in Table 1, the M3 set has the value of q_O that is closest to these values. Thus we shall consider the M3 set to be our primary test set, and the results for the other three sets of charges should be considered just as a way to show the sensitivity or insensitivity to choice of charge set.

In the rest of this article and in all tables, the 6-31+G(d,p) basis will be abbreviated DZ, and the 6-311++G(2d,2p) basis will be abbreviated TZ. In the rest of the article the combination of a density functional and a basis set or of a wave function method and a basis set will be called a level.

We obtain our best estimates of the energy E by a standard type of extrapolation procedure, namely

$$\text{E(extrap)} = \text{QCISD(T)/6-311+G (d,p)} + \text{MP2/TZ} - \text{MP2/6-311+G(d,p)} \quad (1)$$

3. Results and Discussion

3.1. Full Calculations of Tetramers and Pentamers. Tables 2 and 3 show the energies (relative to the ring configuration) of the structures of the tetramers and pentamers used in this paper as calculated at various levels of theory.

The results calculated with DFT are very sensitive to the basis set. In the case of tetramers the energies calculated with the small DZ and 6-311+G(d,p) basis sets are closer to the CCSD(T)/DZ values calculated here and to the QCISD(T)[29]/6-311+G(d,p) results taken from ref 16 than are the energies calculated with the larger TZ basis set.

As was found previously for small water clusters,[1] the results obtained by MP2 and CCSD(T) are in semiquantitative agreement with each other in our calculations. The differences, with the DZ and 6-311+G(d,p) basis sets, do not exceed 0.38 kcal/mol in either Table 2 or Table 3. The energy ordering of tetramers agrees for all the methods except PBE and M06-2X; PBE is the only method that finds structure E to be more favorable than D, and M06-2X predicts structure C to be lower in energy than structure B. The results of the B3LYP and M06-2X methods with the DZ basis set are close to MP2, CCSD(T)/DZ, and QCISD(T)/6-311+G(d,p) for the B and D structures, but for the C structure the results differ by more than 1 kcal/mol.

In the case of pentamers the extrapolated calculation predicts that structure B is the lowest-energy structure, but most nonextrapolated calculations predict that the energy of the ring configuration A is lower. The reason why density functional theory and wave function theory give different trends is unknown. However the main focus of the present study is how well the EE-MB for a given level reproduces a full calculation for a given level, not the accuracies of the individual levels. The EE-MB is considered useful if it can reproduce a full calculation at a given level with a mean deviation smaller than a reasonable expectation of the error in the full calculation.

3.2. EE-MB Results for Absolute Energies of Tetramers and Pentamers. In order to evaluate the usefulness of the electrostatically embedded many-body method in the case of ammonia−water clusters, we compare their predicted energies to the results of the full calculations at each level of theory. Table 4 shows the mean unsigned deviations (MUDs) between the electrostatically embedded two- and three-body calculations for tetramers and pentamers and the full calculations. As was anticipated from studies of pure water clusters[1] the errors of the EE-PA calculations are 5−10 times larger than the errors of EE-3B calculations. Additional tables given in the Supporting Information show that the errors in the pentamer energies are larger than the errors of the tetramers; for tetramers the MUDs with the M3 charge set do not exceed 0.10 kcal/mol per monomer for EE-PA and 0.023 kcal/mol per monomer for EE-3B; for pentamers these MUEs rise to 0.17 kcal/mol per monomer and 0.084 kcal/mol per monomer for EE-PA and EE-3B, respectively.

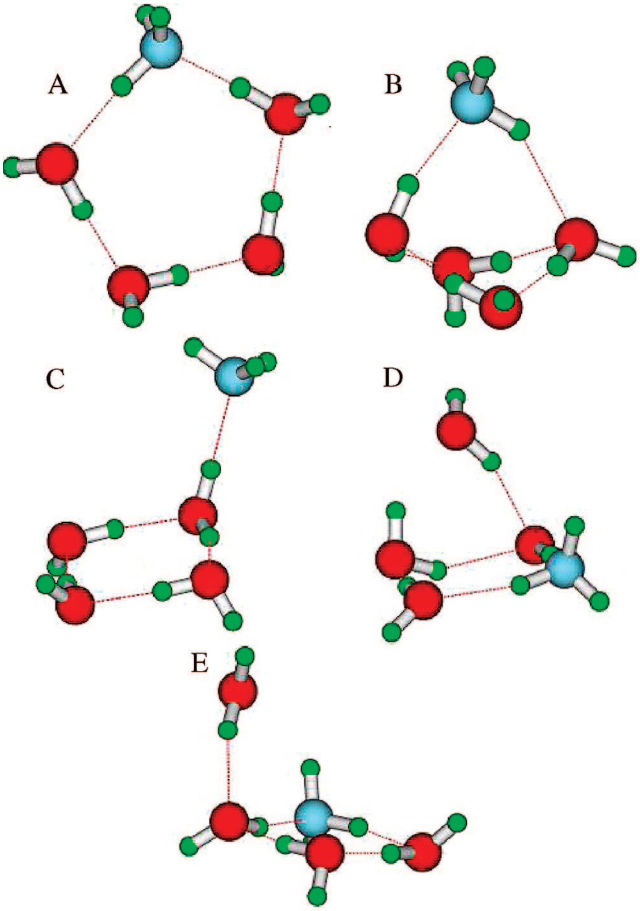


Figure 2. Structural isomers of $NH_3(H_2O)_4$ (pentamers) used in this work.

Almost all results show that the errors of calculations with M1 and CM4 charges are largest, while the calculations with M3 charges are the most accurate. That means that the electrostatically embedded many-body calculations of these mixed clusters are more accurate with larger charges on either kind of monomer. The EE many-body calculations for water clusters show the same trends.[1] It is, however, important to keep the comparison of charge models in context; that is, none of the charge models yields unacceptably large errors. In particular, the largest deviation in the MUD for any two sets of charges (at a given level of theory) never exceeds 0.35 kcal/mol at the EE-PA level of theory or 0.04 kcal/mol at the EE-3B level of theory.

3.3. EE-MB Results for Relative Energies of Tetramers and Pentamers. With five structures, there are ten pairs of structures, and it is interesting to test how well EE-MB can predict these relative energies. Table 5 shows the results for tetramers, and Table 6 shows them for pentamers. Both tables show the mean unsigned deviation of the electrostatically embedded pairwise and three-body calculations of relative energies as compared to full calculations at the same level. The

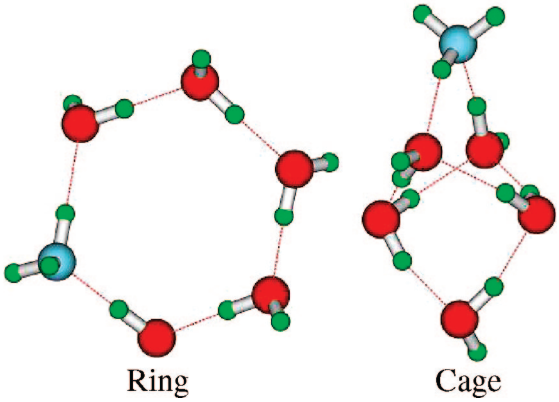


Figure 3. Structural isomers of $NH_3(H_2O)_5$ (hexamers) used in this work.

M06-2X, MP2, and CCSD(T) methods show the best agreement. The magnitudes of the mean unsigned deviations are encouragingly small; with the M3 model charges the mean unsigned deviations are 0.16-0.76 kcal/mol (average: 0.41 kcal/mol) for EE-PA (top halves of Tables 5 and 6) and 0.01−0.31 kcal/mol (average: 0.10 kcal/mol) for EE-3B (lower halves of Tables 5 and 6).

Note that the deviations from the full result due to truncating the expansions at second order (top halves of Tables 5 and 6) are smaller than the typical deviation of the full calculations from the extrapolated ones in Tables 2 and 3, and the deviations due to truncating at third order (lower halves of Tables 5 and 6) are even smaller. Thus the combination of the truncated many-body expansions with DFT and/or MP2 calculations provides economical “model chemistries”[30–33] that should be as useful as the untruncated DFT and MP2 calculations for many applications but at considerably reduced cost for large systems.

3.4. Full Calculations and EE-MB Results for Hexamers. The hexamer is the hardest test of the usefulness of EE-MB because the PA and 3B approximations omit the most interactions for these largest clusters. Our full calculations of the $NH_3(H_2O)_5$ clusters show that the cage structure is more favorable than the ring one. The energies (in kcal/mol) of the cage geometry of the $NH_3(H_2O)_5$ hexamer relative to the ring geometry as predicted by full calculations are −2.02 for PBE/TZ, −1.28 for B3LYP/TZ, −5.69 for M06-2X/TZ, and −2.96 for MP2/TZ. This shows that, of four full calculations, the M06-2X/TZ method predicts the largest energy gap between these two structures.

Table 7 shows that the EE-MB method agrees well with full calculations for the hexamer configurations. For the M3 charges, the MUD per monomer for hexamers does not exceed 0.10 kcal/mol per monomer for the EE-PA method and 0.055 kcal/mol per monomer for the EE-3B method. The results for hexamers show the same charge trend as was observed in tetramers and pentamers, namely that the deviation between the full and EE-MB calculation is smaller when the charges of all atoms in the water and ammonia molecules are larger, as in the M3 model charges. The EE-PA calculations with the M06-2X and MP2 methods show the best agreement with full calculations, while for EE-3B calculations the B3LYP and MP2 show the smallest truncation errors.

4. Conclusions

The present test of the electrostatically embedded many-body method is an important step in its validation because we consider mixed clusters with the same kinds of choices for the charge models that were previously successful for pure water clusters. Furthermore, the polarization of ammonia has new aspects because it is nonplanar. Therefore, it is encouraging that our calculations show that the electrostatically embedded three-body approximation is very accurate for calculations of small ammonia−water clusters (tetramers, pentamers, and hexamers), and the electrostatically embedded pairwise additive approximation also provides useful accuracy. The success for clusters with mixed electrostatics, the good scaling properties of the pairwise and three-body approximations, the fact that the error per monomer does not increase when the size of the cluster increases, and our recent demonstration[34] that the truncated expansions can yield accurate and convenient gradients are all promising features for future applications to large systems.

Acknowledgment. This work was supported in part by the National Science Foundation under grant ITR-04228774.

Supporting Information Available: Coordinates of the clusters and additional error tables (22 tables). This material is available free of charge via the Internet at http://pubs.acs.org.floyd.lib.umn.edu.

References

(1) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.

(2) Komeiji, Y.; Tatsuya, N.; Fukuzawa, K.; Ueno, Y.; Inadomi, Y.; Nemoto, T.; Uebayasi, M.; Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2003**, *372*, 342.

(3) Fedorov, D. G.; Kitaura, K. J. *Phys. Chem. A* **2007**, *111*, 6904.

(4) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. *Mol. Phys.* **2005**, *103*, 2255.

(5) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput*, to be published.

(6) Weber, R. J.; Marti, J. J.; McMurry, P. H.; Eisele, F. L.; Tanner, D. J.; Jefferson, A. *J. Geophys. Res. Atmos.* **1997**, *102*, 4375.

(7) Weber, R. J.; McMurry, P. H.; Mauldin, L.; Tanner, D. J.; Eisele, F. L. *J. Atmos. Sci.* **1995**, *52*, 2242.

(8) Dyke, T. R.; Herbine, P. *J. Chem. Phys.* **1985**, *83*, 3768.

(9) Stockman, P. A.; Bumgarner, R. E.; Suzuki, S.; Blake, G. A. *J. Chem. Phys.* **1992**, *96*, 2496.

(10) Canagaratna, M.; Phillips, J. A.; Ott, M. E.; Leopold, K. R. *J. Phys. Chem. A* **1998**, *102*, 1489.

(11) Eisele, F. L.; Hanson, D. R. *J. Phys. Chem. A* **2002**, *104*, 830.

(12) Chen, B.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 8725.

(13) Astrand, P.-O.; Karlstrom, G.; Engdahl, A.; Nelander, B *J. Chem. Phys.* **1995**, *102*, 3534.

(14) Lee, C.; Fitzgerald, G.; Planas, M.; Novoa, J. J. *J. Phys. Chem.* **1996**, *100*, 7398.

(15) Donaldson, D. J. *J. Phys. Chem. A* **1999**, *103*, 62.

(16) Bacelo, D. E. *J. Phys. Chem. A* **2002**, *106*, 11190.

(17) (a) Zhao, Y.; Truhlar, D. G. *MN-MGFM: Minnesota Gaussian Functional Module - Version 3.0;* University of Minnesota: Minneapolis, MN, 2006. http://comp.chem.umn.edu/mn-gfm/. (b) *Gaussian 03, Revision C.02*; M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. A. Pople, Gaussian, Inc.: Wallingford, CT, 2004.

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

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

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

(21) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(23) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.*, in press. Published online at http://dx.doi.org/10.1007/s00214-007-0310-x.

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

(25) Raghavachari, K.; Anderson, J. B. *Chem. Phys. Lett.* **1989**, *157*, 479.

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

(27) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(28) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

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

(30) Pople, J. A. In *Energy, Structure, and Reactivity*; Smith, D. W., McRae, W. B., Eds.; Wiley: New York, 1973; p 51.

(31) Head-Gordon, M. *J. Phys. Chem.* **1996**, *100*, 13213.

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

(33) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1009.

(34) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1.

CT7003462

JCTC Journal of Chemical Theory and Computation

Visualization of Molecular Orbitals and the Related Electron Densities

Maciej Haranczyk*[†] and Maciej Gutowski*[†,‡]

Department of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland, and Chemistry-School of Engineering and Physical Sciencs, Heriot-Watt University, Edinburgh EH14 4AS, U.K.

Received February 10, 2008

Abstract: When plotting different orbitals with consistent contour values, one can create illusions about the relative extension of charge distributions. We suggest that the comparison is not biased when plots reproduce the same fraction of the total charge. We have developed an algorithm and software that facilitate this type of visualization. We propose superimposing molecules and associated orbitals, and creating cross-sections by selecting a particular part of the orbital limited by pre-defined planes.

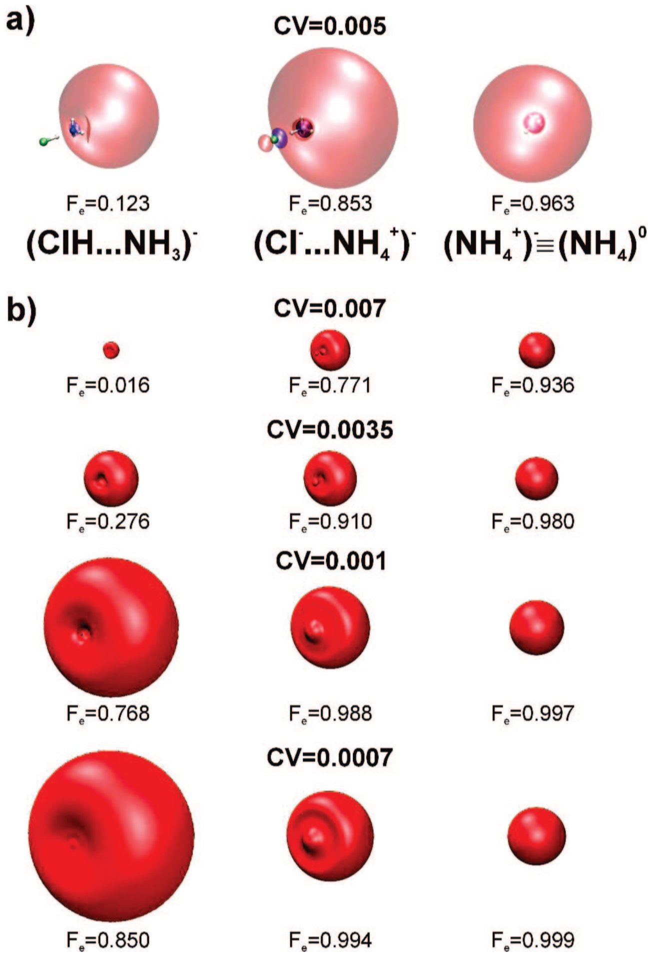


Figure 1. (a) Singly occupied molecular orbitals of the dipole-bound intermediate $(ClH\cdots NH_3)^-$ (left), the proton-transferred species $(Cl^-\cdots NH_4^+)^-$ (center), and, for comparison, the neutral Rydberg radical (NH_4^0) (right). The orbitals were plotted using a contour value (CV) of 0.005 bohr$^{-3/2}$. (b) Positive part of the orbital (pink in part a)) plotted with different CVs (in bohr$^{-3/2}$).

Molecular orbitals and the related electron densities are basic molecular features of interest to chemists. The values of electron density in a molecular fragment and the bonding/antibonding character of the orbital contribute to the chemical properties of this fragment. Therefore, practically all electronic structure codes give their users an option to access molecular orbital data, either in the form of the coefficients associated with basis functions or as volumetric data with values of the orbital or the related electron density at each point of a predefined grid. Many programs have been developed to visualize orbitals and/or electron density, and they are in common use by the community of computational chemists.[1–3] Orbitals and electron densities are typically visualized as finite volumes limited by a boundary defined by a preselected contour value (CV). On occasion, 2D maps, which are cross-sections of the finite volumes, are prepared with marked isovalues of the presented quantity.

Interestingly, plotting an orbital or electron density with a predefined contour value seems to be the only option implemented in the major visualization software packages.[1–3] Similarly, when comparing molecular orbitals or electron densities of different systems, one usually prepares plots using consistent CVs. This approach works well when the charge distributions do not differ much in their spatial extension. We found, however, the same approach to be misleading when the studied charge distributions span a broad range of extension. The problem becomes particularly relevant when dealing with orbitals, which are characterized by very different orbital energies, and therefore different electron binding energies. This results from the long-range asymptotic behavior of bound-state wave functions and orbitals:[4,5] e.g., the occupied Hartree–Fock orbitals decay as exp $[-(-2\varepsilon_{HOMO})^{1/2}r]$,[4] where ε_{HOMO} is the orbital energy of the highest occupied orbital. Here we will focus

* To whom correspondence should be addressed. E-mail: maharan@chem.univ.gda.pl (M.H.); m.gutowski@hw.ac.uk (M.G.).

[†] University of Gdańsk.

[‡] Heriot-Watt University.

10.1021/ct800043a CCC: $40.75
Published on Web 04/05/2008

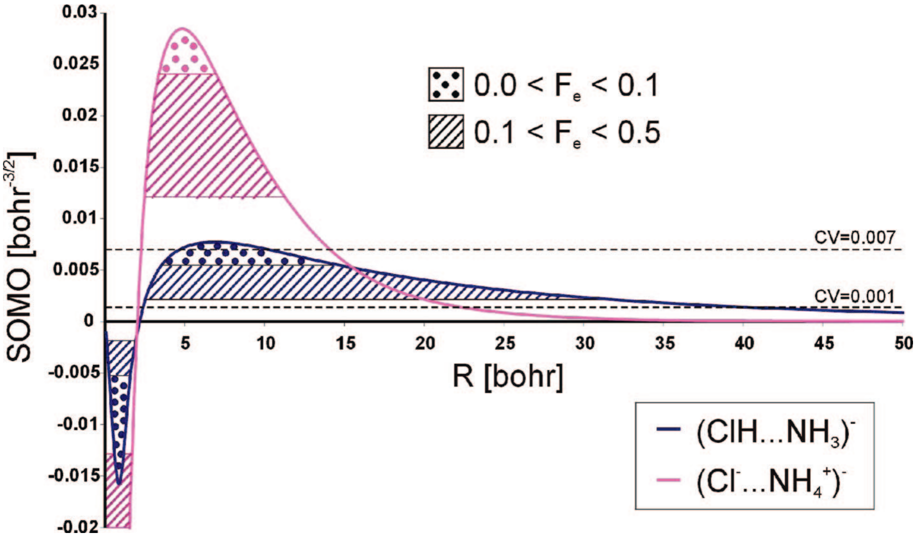


Figure 2. Decay of the SOMO orbitals. *R* defines the distance from N along the N−Cl line in the direction of the main lobes of the SOMOs.

Table 1. Algorithm for Determination of a Contour Value Corresponding to a Preselected Fraction of the Total Orbital Charge

1. generate or read-in grid points and the corresponding volumetric data containing orbital or orbital density values
2. if the orbital values are provided in point 1, calculate the corresponding orbital density values
3. sort grid points according to the orbital density values
4. loop over sorted grid points and perform numerical integration of the orbital density starting from the point of the highest density value
5. stop integration when the integrated value exceeds the preselected fraction
6. the searched CV is equal to the value of orbital density at the last integrated point (if plotting electron densities) or to the properly signed square root of it (if plotting orbitals)

on molecular anions, though the visualization of orbitals for neutral and cationic species encounters similar problems, e.g., when comparing Rydberg and valence orbitals of neutral species.

We will use two examples from our recent studies on molecular anions: an anionic complex of ammonia and hydrogen chloride,[6] and tautomers of an anionic nucleic acid base, guanine.[7–10] The selected systems cover various types of molecular anions: a dipole-bound anion,[6] a related system, in which a closed-shell anion is bound to a neutral Rydberg molecule,[6] and valence anions.[7–10]

The complex of ammonia and hydrogen chloride has been a subject of our recent study.[6] The neutral system has a hydrogen-bonded ($ClH\cdots NH_3$) structure in the gas phase.[11,12] Since it has a dipole moment of ca. 4.15 D, it supports a dipole-bound state with the excess electron attached to NH_3 (Figure 1a, left), and the calculated electron vertical attachment energy is 0.03 eV.[6] The excess electron attachment modifies the potential energy surface of ($ClH\cdots NH_3$) and promotes an intermolecular proton transfer, with the global anionic minimum having the $(Cl^-\cdots NH_4^+)^-$ form (Figure 1a, middle), and the excess electron remains bound to the nitrogen site, but with a much larger binding energy, 0.51 eV.[6] The latter anionic complex can also be characterized as a Rydberg radical, $(NH_4)^0$ (Figure 1a, right), polarized by Cl^-. The isolated, fully symmetric $(NH_4)^0$ radical is characterized by a vertical ionization potential of 5.08 eV.[6]

The significant differences in electron binding energies among $(ClH\cdots NH_3)^-$, $(Cl^-\cdots NH_4^+)^-$, and $(NH_4)^0$ should be reflected in the diffuseness of the singly occupied molecular orbital (SOMO). The corresponding SOMOs are presented in Figure 1a and were prepared according to the common practice—a consistent CV of 0.005 $bohr^{-3/2}$ was selected. Surprisingly, the SOMO of $(Cl^-\cdots NH_4^+)^-$ seems to be more diffuse than that of $(ClH\cdots NH_3)^-$, even though the former is characterized by an electron binding energy 1 order of magnitude larger than the latter. Below we demonstrate how this illusion develops and how to avoid it.

In Figure 1b we plotted bulky parts of the SOMOs using different contour values: 0.007, 0.0035, 0.001, and 0.0007 $bohr^{-3/2}$. For CV = 0.007 $bohr^{-3/2}$ the orbital bulb of $(ClH\cdots NH_3)^-$ is much smaller than for $(Cl^-\cdots NH_4^+)^-$ or $(NH_4)^0$. However, the opposite size relation might be deduced for CV = 0.0007 $bohr^{-3/2}$. Clearly, the visualization of molecular orbitals with the same CV value might fail to provide information about their relative sizes.

An important observation is that the fractions of electrons (F_e) contained in the volumes determined by the same CV value might be very different (Figure 1). For example, the SOMOs presented in Figure 1a reproduce 12.3% of e for $(ClH\cdots NH_3)^-$, 85.3% for $(Cl^-\cdots NH_4^+)^-$ and 96.3% for NH_4^0. Moreover, it requires quite a small CV value to reproduce a significant fraction of e for $(ClH\cdots NH_3)^-$. We conclude that the inability to derive information about the relative orbital sizes is related to the inconsistent F_e values.

To illustrate the point further, we show in Figure 2 plots of the SOMO for $(ClH\cdots NH_3)^-$ and $(Cl^-\cdots NH_4^+)^-$ as a function of *R*, where *R* is the distance from the nitrogen atom along the Cl−N line in the direction of the main lobes. The SOMO of $(Cl^-\cdots NH_4^+)^-$ decays much faster than that for $(ClH\cdots NH_3)^-$, as anticipated from the electron binding energies.[4] Both orbitals are normalized to 1, and therefore the maximum value of SOMO for $(ClH\cdots NH_3)^-$ must be smaller than for $(Cl^-\cdots NH_4^+)^-$. Consequently, if the CV is sufficiently large, then the criterion SOMO(R) > CV is met only for small values of *R* for $(ClH\cdots NH_3)^-$ but for larger *R* values for $(Cl^-\cdots NH_4^+)^-$. This

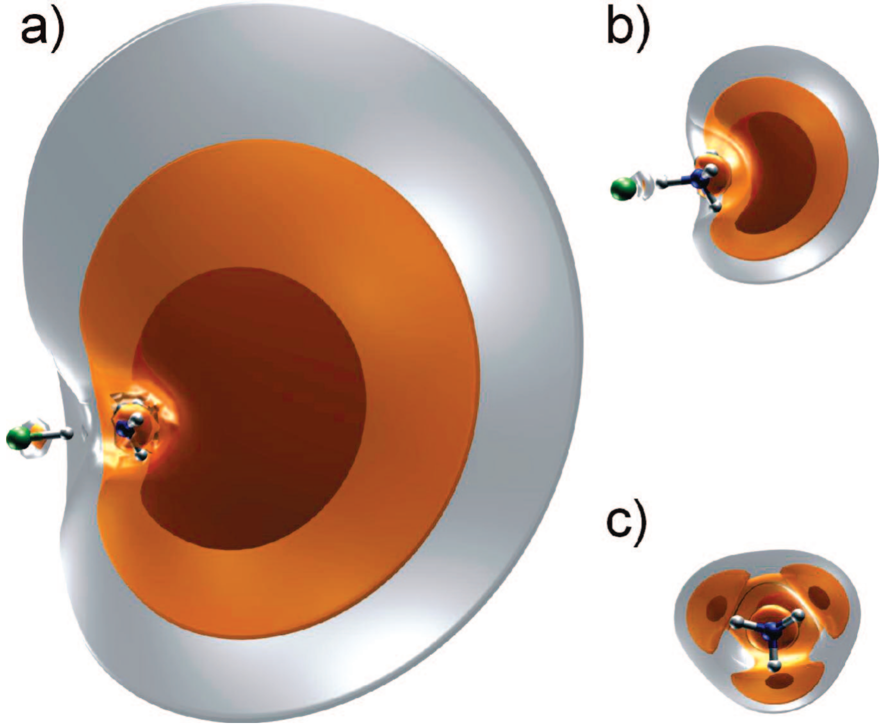


Figure 3. Cross-sections through the density of the SOMO for (a) the dipole-bound intermediate $(ClH\cdots NH_3)^-$, (b) the proton-transferred species $(Cl^-\cdots NH_4^+)^-$, and for comparison, (c) the neutral Rydberg radical (NH_4^0). These plots were generated with VMD[1] and OpenCubMan,[16] and the resulting contours enclose 0.1, 0.3 and 0.5 e from the inner to the outer shell, respectively.

explains why the plotted orbital is larger for $(Cl^-\cdots NH_4^+)^-$ than for $(ClH\cdots NH_3)^-$ for CV = 0.007 and 0.005 bohr$^{-3/2}$ (Figure 1). We re-emphasize that in these cases only small fractions of *e* are reproduced for $(ClH\cdots NH_3)^-$, whereas the F_e values exceed 0.7 e for $(Cl^-\cdots NH_4^+)^-$. It requires quite a small value of CV to have the SOMO of $(ClH\cdots NH_3)^-$ described by a lobe larger than that for $(Cl^-\cdots NH_4^+)^-$,: see the cases of CV = 0.001 and 0.0007 bohr$^{-3/2}$ in Figures 1 and 2. In these cases the F_e values exceed 0.7 e for all three systems.

In Figure 2 we also visualize these ranges of *R* that have to be included to reproduce a given value of F_e, and the two illustrated cases are $F_e = 0.1$ and 0.5 e. If the orbital plots were based on the criterion of the same value of F_e, then the SOMO of $(ClH\cdots NH_3)^-$ would be more diffuse than that for $(Cl^-\cdots NH_4^+)^-$. This suggests that an unbiased way to visualize orbitals or electron densities that differ much in the extension of charge distributions would be to ensure that a consistent and preselected fraction of the total charge is reproduced in each plot. The same conclusion was reached by Rauk and Armstrong in their studies of dipole-bound and valence anions in clusters involving various hydrogen halides.[13–15] The approach, i.e., plotting different orbitals in such a way that the same fraction of electron charge is reproduced, leads to another question: what are the CV's that lead to the same and preselected F_e's? Clearly, these CV's might be different for different orbitals. Here we present an efficient algorithm to determine the desirable CV's. The same algorithm can be used to calculate a fraction of the total charge corresponding to a particular CV. We also suggest how to graphically present information about the relative diffuseness of orbitals. First we make a list of a few F_e's, and we create all the 3D orbital plots for the preselected F_e's. Finally, for each orbital we superimpose plots corresponding to the preselected F_e's and we create 2D cross-sections that unravel information about the relative orbital diffuseness.

The proposed approaches are made available to the scientific community by providing appropriate software.[16] This software works with volumetric data containing orbitals or orbital densities. The latter are often referred to as "cube files".[17] They typically contain the Cartesian coordinates of atoms and a definition of the grid. The grid is defined by a starting point, three nonparallel vectors, and the size of the grid (the numbers of points in each direction defined by the grid vectors). Our software provides the following functionality: (i) identification of a CV that corresponds to a preselected value of F_e, (ii) determination of F_e associated with a given CV, and (iii) selection of a particular part of the grid limited by a predefined plane. This selection is made by zeroing the to-be-discarded part of the grid. The last functionality can be applied many times: i.e., a few planes can be defined and the grid can be trimmed to the desired slice of the orbital or the related electron density. It is up to the user to define desirable F_e's and limiting planes, if any. We believe that instructive plots of orbitals and orbital densities can be generated using the OpenCubMan software[16] in combination with molecular visualization packages and using "cube files" produced by common quantum chemistry packages.

A CV corresponding to a preselected F_e is determined using the algorithm summarized in Table 1. In this algorithm the charge density is integrated by starting from the most dense region to the least dense region. The process of density integration is stopped when a preselected fraction of the charge has been recovered. The searched CV is equal to the value of

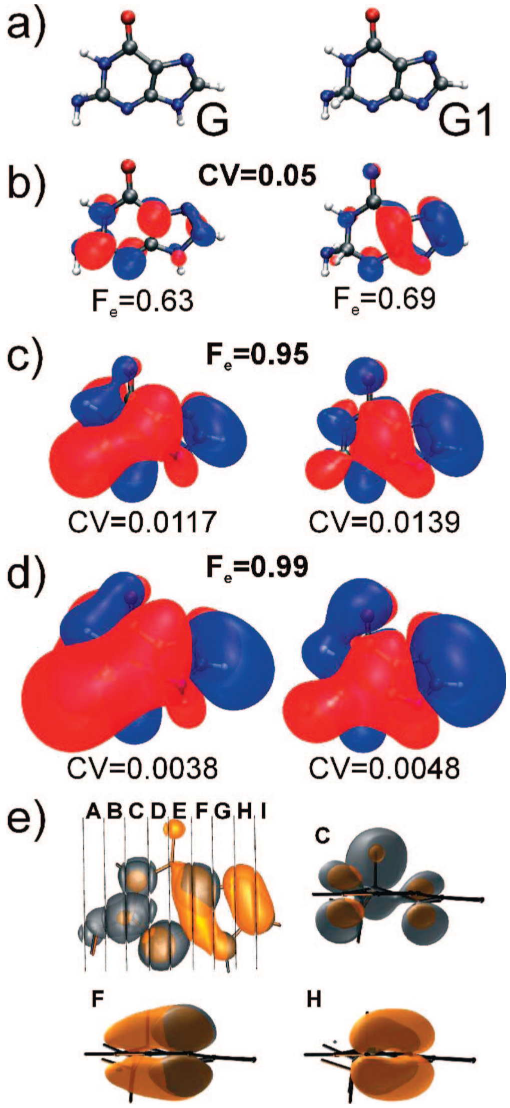


Figure 4. (a) Molecular structures of valence anions of the canonical tautomer of guanine (G) and the most stable anionic tautomer (G1). (b) Singly occupied molecular orbitals plotted using a contour value of 0.05 bohr$^{-3/2}$. (c) and (d) Singly occupied molecular orbitals plotted with F_e equal to 0.95 and 0.99, respectively. (e) Selected cross-sections of singly occupied molecular orbital densities corresponding to 0.6 e. The SOMO densities for G and G1 are superimposed and distinguished with gray and yellow, respectively. All cross-sections, marked A−I, are available in the Supporting Information (Figure S-1).

orbital density at the last integrated point (if plotting electron densities) or to the properly signed square root of it (if plotting orbitals).

Creating a cross-section of an orbital represented on a grid can be achieved by zeroing a part of the volumetric data above or below a predefined plane. A given plane is described as

$$ax + by + cz + d = 0 \qquad (1)$$

where a, b, and c are components of a vector $\mathbf{v}$ normal to the plane and d is a parameter which can be calculated by solving eq 1 for a given point on the plane. A distance D of any point $p_0 = (x_0,y_0,z_0)$ from the plane can be calculated using the equation[18]

$$D = \frac{ax_0 + by_0 + cz_0 + d}{\sqrt{a^2 + b^2 + c^2}} \qquad (2)$$

Such a definition allows D to have a positive or negative sign. D is positive if p_0 is on the same side of the plane as the vector $\mathbf{v}$ and negative if it is on the opposite side. When a part of the grid is zeroed by using a plane, each point of the grid is tested against eq 2, and the value of this point is set to zero or remains unchanged, if appropriate.

All of the functions presented above have been implemented in the Open-Source Cubefile Manipulator (OpenCubMan) program, which is provided free of charge under the GNU license[16] and can be downloaded from the SourceForge Internet Archive. OpenCubMan was written in the object oriented C++ programming language and is provided as a C++ object definition. OpenCubMan uses standard C/C++ libraries for all input/output operations, math, and sorting (qsort function). This form facilitates incorporating the code into other packages, libraries, or scripting languages.

The orbitals and orbital densities considered here and presented in Figures 1, 3, and 4 are based on UHF calculations for the $(\mathrm{ClH}\cdots\mathrm{NH_3})^-$, $(\mathrm{Cl^-}\cdots\mathrm{NH_4^+})^-$, and $\mathrm{NH_4^0}$ systems as well as for the valence anions of guanine. The aug-cc-pVDZ basis set was used for all systems, with additional diffuse functions for the first three systems.[6] The “cube files” with SOMO orbitals were prepared using the Gaussian03 program.[17] They were later modified with the OpenCubMan program[16] and then visualized using the VMD software.[1]

OpenCubMan was used to determine: (i) the values of F_e corresponding to preselected CVs (Figure 1) and (ii) the values of CV corresponding to preselected values of $\mathrm{F_e}$ (Figures 3 and 4). The most time-consuming part of the algorithm (Table 1) is the third step, in which the values of electron density on the grid are sorted. We typically used a consistent grid of 250 × 180 × 180 points containing volumetric data for the SOMO orbitals. Sorting the electron density values on this grid took about 2 s on the available Intel Pentium 4 computer. The algorithm was also tested for a larger grid of 64 000 000 points (400 × 400 × 400) generated for the most diffuse SOMO of $(\mathrm{ClH}\cdots\mathrm{NH_3})^-$. The sorting time exceeded 100 s, which is not significant in comparison with the time required to generate such a dense grid. Moreover, such extended grids are used only in very special cases, such as dipole-bound anions with very small electron binding energies. Therefore, they will not be used in typical applications.

Next, we visualized the SOMO electron densities for $(\mathrm{ClH}\cdots\mathrm{NH_3})^-$, $(\mathrm{Cl^-}\cdots\mathrm{NH_4^+})^-$, and $\mathrm{NH_4^0}$ and three preselected values of F_e, 0.1, 0.3 and 0.5 e, and the results are presented in Figure 3. For each system we superimposed plots corresponding to the three F_e values and we created 2D cross-sections that unravel information about the relative diffuseness of the SOMO distributions. The cross-sections were produced with OpenCubMan using an approach described above. The thickness of the consecutive layers is the largest for $(\mathrm{ClH}\cdots\mathrm{NH_3})^-$ and the smallest for $\mathrm{NH_4^0}$, thus unraveling that the SOMO of the former is the most diffuse and that of the latter the least diffuse.

Finally we applied OpenCubMan to visualize SOMOs of valence anions of guanine. We selected the canonical tautomer (G) and the most stable anionic tautomer (G1), which we have studied in the past,[7–10] and their structures are shown in Figure

4a. When the SOMOs of G^- and $G1^-$ are visualized with the same CVs of 0.05 bohr$^{-3/2}$ (Figure 4b), then the corresponding F_e's are 0.629 and 0.694 e. Clearly, the 6.5% difference is significant, though not as large as in the $(ClH\cdots NH_3)^-$ and $(Cl^-\cdots NH_4^+)^-$ systems. The calculated electron vertical detachment energies are 0.59 and 2.43 eV or G^- and $G1^-$, respectively. We selected consistent values of F_e of 0.95 and 0.99 e, and the resulting SOMOs are shown in parts c and d of Figure 4, respectively. The plots illustrate the different bonding/antibonding character of these orbitals, which leads to different values of VDE.[19] The SOMO is more diffuse for G^- than for $G1^-$, though the differences are much smaller than for the $(ClH\cdots NH_3)^-$ and $(Cl^-\cdots NH_4^+)^-$ systems.

Finally, we show a plot that illustrates differences in the spatial distribution of the excess electron in G^- and $G1^-$. In Figure 4e and Figure S-1 (Supporting Information) we superimpose both tautomers and the corresponding SOMOs and we focus attention on nine slices, which are selected by applying specific planes. For G^-, the majority of the excess electron in localized on the six-membered ring, whereas for $G1^-$ the excess electron is localized on the five-membered ring.

In conclusion, we developed a capability to visualize molecular orbitals that differ much in the extension of charge distribution. We recommend that these plots should reproduce the same fraction of the total charge to avoid illusions that develop when constructing plots with the same contour values. The OpenCubMan software facilitates operations on common "cube files", allows superimposing molecules and associated orbitals, and selects a particular part of the orbital limited by predefined planes.

Acknowledgment. Helpful discussions with Rafał A. Bachorz and Soren Eustis are gratefully acknowledged. This work was supported by the Polish State Committee for Scientific Research Grants Nos. DS/8221-4-0140-8 (to M.G.) and N204 127 31/2963 (to M.H.). M.H. holds an award from Foundation for the Development of the University of Gdańsk (FRUG). The calculations were performed at the Academic Computer Center in Gdańsk (TASK) and at the Molecular Science Computing Facility (MSCF) in the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL), where computer time was provided via a Computational Grand Challenge Project.

Supporting Information Available: Figure S-1 presenting the A—I cross sections of the SOMO densities for G^- and $G1^-$ tautomers. This material is available free of charge via the Internet at http://pubs.acs.org.

References

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

(2) Schaftenaar, G.; Noordik, J. H. *J. Comput.-Aided Mol. Design* **2000**, *14*, 123–134.

(3) Black, G.; Didier, B.; Bisethagen, T.; Feller, D.; Gracio, D.; Hackler, M.; Havre, S.; Jones, D.; Jurrus, E.; Keller, T.; Lansing, C.; Matsumoto, S.; Palmer, B.; Peterson, M.; Schuchardt, K.; Stephan, E.; Sun, L.; Taylor, H.; Thomas, G.; Vorpagel, E.; Windus, T.;Winters, C., Ecce, A Problem Solving Environment for Computational Chemistry, Software Version 3.2.5; Pacific Northwest National Laboratory, Richland, WA 99352–0999, 2006.

(4) Handy, N. C.; Marron, M. T.; Silverstone, H. J. *Phys. Rev.* **1969**, *180*, 45–48.

(5) Katries, J.; Davidson, E. R. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 4403–4406.

(6) Eustis, S. N.; Radisic, D.; Bowen, K. H.; Bachorz, R. A.; Haranczyk, M.; Schenter, G. K.; Gutowski, M. *Science* **2008**, *319*, 936–939.

(7) Haranczyk, M.; Gutowski, M. *Angew. Chem., Int. Ed.* **2005**, *44*, 6585–6588.

(8) Haranczyk, M.; Gutowski, M. *J. Chem. Inf. Model.* **2007**, *47*, 686–694.

(9) Haranczyk, M.; Gutowski, M. *J. Am. Chem. Soc.* **2005**, *127*, 699–706.

(10) Haranczyk, M.; Gutowski, M.; Li, X.; Bowen, K. H. *J. Phys. Chem. B* **2007**, *111*, 14073–14076.

(11) Howard, N. W.; Legon, A. C. *J. Chem. Phys.* **1988**, *88*, 4694.

(12) Legon, A. C. *Chem. Soc. Rev.* **1993**, *22*, 153–164.

(13) Rauk, A.; Armstrong, D. A. *Int. J. Quantum Chem.* **2003**, *95*, 683–696.

(14) Li, X.; Sanche, L.; Rauk, A.; Armstrong, D. A. *J. Phys. Chem. A* **2005**, *109*, 4591–4600.

(15) Rauk, A.; Armstrong, D. A. *Eur. Phys. J. D* **2005**, *35*, 217–224.

(16) Open-source Cubefile Manipulator Program (OpenCubMan) is available free of charge at the SourceForge archive: http://opencubman.sourceforge.net (accessed Feb 27, 2008).

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

(18) Gellert, W.; Gottwald, S.; Hellwich, M.; Kästner, H.; Künstner, H., Eds. *VNR Concise Encyclopedia of Mathematics*, 2nd ed.; Van Nostrand Reinhold: New York, 1989.

(19) Haranczyk, M.; Holliday, J.; Willett, P.; Gutowski, M. *J. Comput. Chem.*, in press (doi: 10.1002/jcc.20886).

CT800043A

JCTC Journal of Chemical Theory and Computation

Cholesky Decomposition-Based Multiconfiguration Second-Order Perturbation Theory (CD-CASPT2): Application to the Spin-State Energetics of Co^{III}(diiminato)(NPh)

Francesco Aquilante,[‡] Per-Åke Malmqvist,[‡] Thomas Bondo Pedersen,[†,‡] Abhik Ghosh,[§] and Björn Olof Roos*[,‡]

Department of Theoretical Chemistry, Chemical Center, University of Lund, P.O. Box 124, S-221 00 Lund, Sweden, and Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway

Received October 15, 2007

Abstract: The electronic structure and low-lying electronic states of a Co^{III}(diiminato)(NPh) complex have been studied using multiconfigurational wave function theory (CASSCF/CASPT2). The results have been compared to those obtained with density functional theory. The best agreement with ab initio results is obtained with a modified B3LYP functional containing a reduced amount (15%) of Hartree–Fock exchange. A relativistic basis set with 869 functions has been employed in the most extensive ab initio calculations, where a Cholesky decomposition technique was used to overcome problems arising from the large size of the two-electron integral matrix. It is shown that this approximation reproduces results obtained with the full integral set to a high accuracy, thus opening the possibility to use this approach to perform multiconfigurational wave-function-based quantum chemistry on much larger systems relative to what has been possible until now.

1. Introduction

Seen from an inorganic or bioinorganic vantage point, high-level ab initio methods are somewhat of a succès d'estime[1] (for reviews on inorganic and bioinorganic applications of the present type of ab initio calculations, see refs 2–4). By and large, applications of such methods have been limited to small systems that are of limited interest in inorganic and bioinorganic chemistry. Of course, this does not mean that these methods have not been useful at all. Thus, high-level ab initio calculations have been recently deployed to analyze the electronic structures of multiply bonded transition metal and actinide dimer complexes.[5] However, it is density functional theory (DFT) that, in spite of its limited accuracy, has emerged as the standard method for modeling complex processes involving transition metals such as metalloenzyme mechanisms.[6–9] The major reason underlying this state of affairs is of course that much more computational effort is required with the application of the wave-function-based ab initio methods. Limitations in terms of the number of atoms and the size of the basis sets are more severe than they are in DFT. However, the situation vis-à-vis ab initio methods is now changing with the development of much more efficient techniques to treat the basis set problem.

Here, we report an extension of multiconfiguration reference, second-order perturbation theory (CASSCF/CASPT2),[10–12] based on a Cholesky decomposition (CD) of the electron repulsion integral matrix,[13–16] which is considerably faster than earlier implementations of the same method. Application of the Cholesky decomposition approach to electronic structure calculations is not new,[17–21] but only very recently[22] has the approach been successfully extended to multiconfigurational wave function models, such as the popular CASSCF method.[10] The general applicability of the CASSCF

* Corresponding author e-mail: Bjorn.Roos@teokem.lu.se.

[†] Present address: Atomistix A/S, c/o Niels Bohr Institute, Rockefeller Complex, Juliane Maries Vej 30, DK-2100 Copenhagen, Denmark.

[‡] University of Lund.

[§] University of Tromsø.

10.1021/ct700263h CCC: $40.75
Published on Web 04/16/2008

wave function, combined with the accuracy of the CASPT2 correction, affords a unique protocol for unraveling the subtleties of chemical bonds in transition metal systems. The computational expediency afforded by the Cholesky decomposition approach should go a long way toward establishing the CASSCF/CASPT2 method as a valuable tool in the theoretical inorganic and bioinorganic chemist's toolbox.

The Cholesky decomposition-based CASPT2 (CD-CASPT2) method is illustrated here with calculations on the spin-state energetics of the low-coordinate imido complex Co^{III}(diiminato)(NPh). This complex may be regarded as a slightly simplified C_{2v} model of the closely related, diamagnetic complex Co^{III}(nacnac)(NAd) (nacnac = anion of 2,4-bis(2,6-dimethylphenylimido)pentane, Ad = 1-adamantyl), which has been synthesized and structurally characterized.[23] The diamagnetism of this species, which has been attributed to a $(3d_{z^2})^2(3d_{x^2-y^2})^2(3d_{xy})^2$ electronic configuration on the basis of DFT calculations (where the z direction is identified with the Co−N_{imido} axis), is noteworthy in that the σ^* d_{z^2}-based molecular orbital (MO) is occupied preferentially, whereas the corresponding d_{xz}- and d_{yz}-based π^* MOs are left unoccupied.[24] A detailed DFT study,[24] including calculations on the corresponding oxo complex, showed that this electronic configuration may be attributed to both the low-coordinate nature of the metal center and the nature of the imido ligand. However, the DFT studies, where the newer OPTX-based functionals OLYP and OPBE appeared to be the most reliable,[25] also suggested that there should be low-lying paramagnetic excited states. Indeed, another Co^{III}−imido complex with an $S = 0$ ground state, Co^{III}-($Tp^{tBu,Me}$)(NAd) ($Tp^{tBu,Me}$ = hydrotris(3-t-butyl-5-methylpyrazol-1-yl)borate), has been found to exhibit spin-crossover behavior, as a result of the existence of one or more low-lying, paramagnetic excited states.[26,27] Once again, DFT calculations, especially the OLYP and OPBE functionals, appeared to nicely capture the experimental scenario.[28]

That said, the calculation of the spin-state energetics of open-shell transition metal complexes has long been recognized as a difficult problem for DFT.[4] No one functional appears to perform well for all the problematic cases. A number of studies comparing the performance of different functionals vis-à-vis transition metal spin-state energetics have underscored this problem.[29–40]

Accordingly, the calibration of DFT spin-state energetics against high-level ab initio methods is an important goal for current quantum chemistry method development efforts. In this study, we shall compare DFT and CASPT2 results on the vertical spin-state energetics of Co^{III}(diiminato)(NPh). We shall also use this model complex to calibrate the CD-CASPT2 method against the conventional, full integral-based calculations. Both the CD and conventional CASPT2 calculations were performed with a relativistic double-ζ plus polarization (VDZP) basis set, whereas our best results were obtained using a valence triple-ζ plus polarization (VTZP) basis set where only CD-CASPT2 calculations proved feasible. Given the methodological focus of this study, we will not provide a comprehensive list of references to the transition metal imido literature, but instead will refer the reader to a recent review and references therein.[41]

2. Methodology

The present study has been carried out in part using the Cholesky decomposition representation[13,15] of electron repulsion integrals in all stages of the calculations (integrals, self-consistent field (SCF), CASSCF, and CASPT2). A brief review of this approach is given below. Details of the calculations, the choice of basis set, the active orbitals, and so forth, are given below. For more details about the Cholesky decomposition techniques, see the references given.

2.1. Cholesky Representation of the Electron Repulsion Integrals. The computational complexity of the CASPT2 method[11] depends on two parameters: the size of the space spanned by the complete active space (CAS) reference function and the size of the atomic basis set. The former is determined by the choice of the active space and grows factorially with the size of the active space. At present, CASPT2 calculations are therefore restricted to active spaces of about 14–16 orbitals. Fortunately, it turns out that suitable active spaces can be devised even for extended systems, while ensuring a computationally feasible size of the resulting CAS expansion.

On the other hand, the convergence of the results with respect to the size of the atomic basis set can be slow, as in other ab initio methods. When large basis sets are employed, the computational bottleneck of the CASPT2 method lies in the transformation of electron repulsion integrals (ERIs) from an atomic orbital (AO) to an MO basis. Along with the final MO integrals, the AO integrals and all partially transformed integral intermediates are responsible for the significant storage demands of the CASPT2 method.

A strategy for avoiding the expensive MO transformation of the entire ERI matrix and the storage of the AO ERIs derives from the fact that $1/r_{12}$ is a positive definite operator with eigenvalues clustered toward zero. This property allows for a compact representation of the ERIs by means of an incomplete CD of the matrix. The Cholesky representation of AO electron repulsion integrals may be written as[13,15]

$$(\mu\nu|\lambda\sigma) = \sum_{J}^{M} L_{\mu\nu}^{J} L_{\lambda\sigma}^{J} \qquad (1)$$

where Greek indices denote AOs and $L_{\mu\nu}^{J}$ is the Jth Cholesky vector obtained from the matrix decomposition. Due to near linear dependence in the product space of the AOs, the number of vectors M needed to numerically represent the integrals to an accuracy suited for quantum chemical calculations is significantly smaller than the full dimension of the integral matrix. For most applications, the CD usually needs to be converged to an accuracy (δ) no better than 10^{-4}, and the resulting value of M is only about 3–4 times the number of atomic orbital basis functions (N). Notwithstanding the very few applications since its first appearance,[14,42,43] the CD approach has gained interest in recent years as a possible means to speed up correlated calculations.[13,17–20,44]

Recasting the equations for the evaluation of Fock matrices and for the MO transformation of the ERIs directly in terms of Cholesky vectors results in an immediate reduction in the computational costs of most ab initio or DFT methods.[15–17] Simultaneously, there is also an enormous reduction in storage demands compared to conventional integral calculations. For the standard choice of the CD threshold ($\delta = 10^{-4}$), the disk space required to store the AO basis Cholesky vectors is usually about 1–5% of the total size of the AO ERI matrix. The onset of input−output bottlenecks related to the manipulation of these arrays is therefore shifted to significantly larger atomic basis sets, when using the CD approximation. The implementation of the CD-CASPT2 method in this work aims at taking advantage of this particular aspect of the CD representation of the AO ERIs. The Fock matrix needed in CASPT2 can be computed from the AO Cholesky vectors in an efficient way by employing the recently proposed "local K" screening[16] of the exchange contributions.

A more demanding task is the generation of the right-hand side (RHS) of the equation system, which determines the excitation amplitudes. These are directly dependent on MO ERI elements of the type $(pi|qk)$, where p and q are either active or secondary orbitals, while i and k are either inactive or active orbitals. Frozen orbitals are of course excluded. First, transformed Cholesky vectors L^J_{pi} are computed. This task scales as ON^2M, where O is the number of inactive and active orbitals. Subsets of integrals are computed as

$$(pi|qk) = \sum_J^M L^J_{pi} L^J_{qk} \tag{2}$$

and the resulting RHS elements are stored on disk. The assembly of the $(pi|qk)$ integrals scales as O^2V^2M, rather than the ON^4 required by a conventional MO transformation of the ERIs. The gain comes from the reduced prefactor ($O \ll N$), although the overall scaling is still fifth-order.

The present implementation is not yet optimal. The generation of the RHS elements requires an excessive amount of reading and writing to disk, and in spite of the out-of-core handling, a large amount of memory is needed. In future implementations, this part of the calculation will be more open-ended in the sense of requiring much less memory, and the amount of reading and writing will be reduced. We shall return to the subject of an optimal implementation of CD-CASPT2 in future publications. Nonetheless, we wish to stress that the present implementation, although of limited applicability to large systems, allows us to perform CASPT2 calculations that would be impossible with the conventional implementation. This is achieved because the CD method completely bypasses the AO ERIs' storage bottleneck and also because it produces the needed MO integrals at reduced computational costs and input−output overheads. In the present study, full CASSCF/CASPT2 calculations have been performed with a basis set consisting of 869 basis functions. The same CASSCF calculation is an order of magnitude faster in the Cholesky formulation. A more detailed account of the application of the present scheme to CASSCF wave functions has recently been given.[22]

2.2. Details of the Calculations. The Co^{III}(diiminato)(NPh) model complex used in this study is depicted in Figure 1. The molecule is oriented such that the Co is at the origin, the imido nitrogen is on the z axis, and the phenylimido and 1,3-propanediiminato groups are in the xz plane. The planes of the terminal phenyl groups on the diiminato ligand are perpendicular to the plane of the remainder of the molecule. This 43-atom C_{2v} model complex may be viewed as a simplified version of the experimentally studied complex, Co^{III}(nacnac)(NAd).[23] The $S = 0$ ground state of the model complex was optimized with DFT (PW91 and OLYP/STO-TZP) using the ADF-2006 program system[45] and a Slater-type triple-ζ plus double polarization basis set. The structures so obtained were used in all additional DFT and CASSCF/CASPT2 calculations (which provided, in effect, vertical excitation energies). As detailed later, the Co−N$_{\text{imido}}$ distance was also varied in some limited optimization studies at the CASPT2 level of theory.

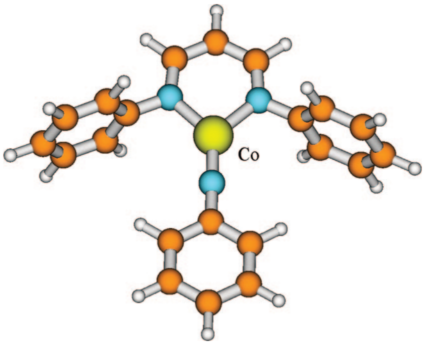
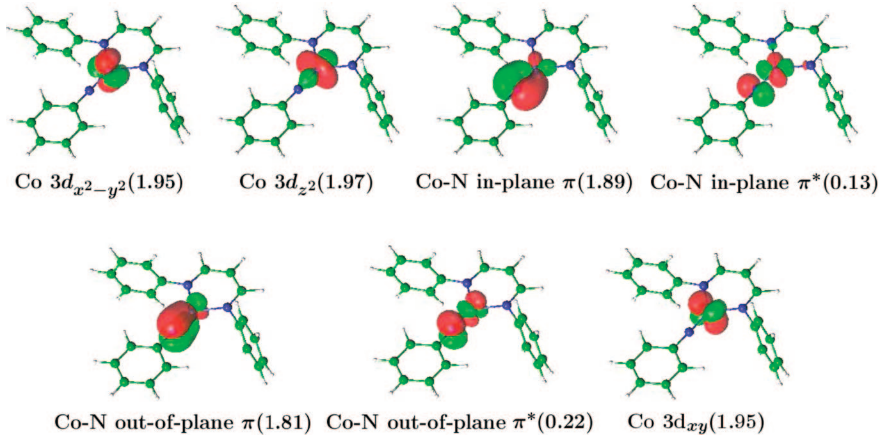


Figure 1. Model of the Co−imido complex used in the calculations.

Two different basis sets were employed in the CASSCF/CASPT2 calculations, both of which were generally contracted atomic natural orbital basis sets including scalar relativistic effects (ANO-RCC).[46,47] This implies that scalar relativistic effects are included at all levels of theory. The smaller one was of VDZP quality: Co/5s4p2d1f, N,C/3s2p1d, and H/2s. With this basis set, we performed calculations using both conventional integrals and the Cholesky approximation with two thresholds, $\delta = 10^{-4}$ and 10^{-8}. The overall basis set consists of 406 basis functions. The larger basis set was of VTZP quality (except for the hydrogens): Co/6s5p3d2f1g, N,C/4s3p2d1f, and H/2s1p. For this basis set, the total number of basis functions is 869. Only Cholesky-based calculations were performed with this basis set. With a threshold of $\delta = 10^{-4}$, it took 3.5 h (wall-clock time on a single AMD Opteron 148, 2.2 GHz, equipped with 1 GB of memory) to generate the Cholesky vectors. The corresponding time with the smaller basis set was 56 min, while the calculation of the full integral set took 151 min. A new method was used for the generation of starting orbitals for this set of calculations. An SCF calculation was performed, and subsequently, the occupied and virtual orbitals of each symmetry were separately localized using the recently developed Cholesky localization technique.[48] The localization, especially for the virtual orbitals, considerably simplified the selection of the physically appropriate orbitals for the active space.

Different active spaces were investigated for the CASSCF calculations, and the final choice was to include the five Co 3d orbitals and the two π orbitals (π_x and π_y) of the imido nitrogen in the active space. Three 4d orbitals were added to account for the "double shell" effect for the doubly



Co $3d_{x^2-y^2}$(1.95) Co $3d_{z^2}$(1.97) Co-N in-plane π(1.89) Co-N in-plane π^*(0.13)

Co-N out-of-plane π(1.81) Co-N out-of-plane π^*(0.22) Co $3d_{xy}$(1.95)

Figure 2. The active molecular orbitals in the Co^{III}(diiminato)(NPh) complex (except the 4d orbitals). A 0.05 au^{-3} level set surface was employed. Occupation numbers are shown in parentheses.

occupied 3d orbitals ($3d_{z^2}$, $3d_{yz}$, and $3d_{x^2-y^2}$).[49] This gives 10 active electrons in 10 orbitals (10in10). The CASSCF natural orbitals for the $^1A_1$ ground state are depicted in Figure 2 (the three weakly occupied 4d orbitals are not shown). Some test calculations were performed with a larger active space, which included also the lone-pair orbitals of the diiminato nitrogens. Their occupation numbers were, however, very close to two, which shows that they are not needed in the active space as long as the electronic states of interest do not involve charge-transfer excitations. Some of the calculations have been performed for a single root in each symmetry and spin. When more than one root was needed, state-average CASSCF calculations were performed.

All 132 valence electrons plus the Co 3s and 3p electrons were correlated in the CASPT2 calculations, which used the standard IPEA Hamiltonian and an imaginary level shift of 0.1 to remove some weak intruder states. The calculations were performed with the MOLCAS-7 quantum chemistry software.[50]

The chemical issue that we sought to address in this project concerns the spin-state energetics of the Co^{III}−imido complex; in other words, how much higher are the $S = 1$ and $S = 2$ states relative to the (experimentally observed) diamagnetic ground state? Accordingly, CASSCF/CASPT2 calculations were performed for the two lowest singlet, triplet, and quintet states in each of the four irreducible representations—in all, 24 electronic states. It turns out that they all have energies below 4 eV, relative to the $^1A_1$ ground state, and it is possible that additional low-lying states would have been found in this energy interval, had the number of roots been extended further. However, for purposes of determining several (∼10) of the lowest electronic states and for a comparison of CASPT2 and DFT energetics, we believe that we have calculated a sufficient number of electronic states.

Density functional calculations were performed for the first state in each symmetry and spin. In some cases, a second state was also computed by locking the number of occupied orbitals in each symmetry. These calculations were performed using a variety of exchange-correlation functionals and a Slater-type triple-ζ plus double polarization basis set. A C_{2v} symmetry constraint was used. The DFT calculations were performed with the ADF program.[45]

Table 1. The Lowest Excited Triplet and Quintet States in Co^{III}(diiminato)(NPh) (Energies in eV) Using a PW91 Ground-State Geometry

state	configuration	energy[a]
$^1A_1$	$(3d_{z^2})^2(3d_{x^2-y^2})^2(3d_{yz})^2(b_1)^2(b_2)^2$	
$^3B_2$	$3d_{x^2-y^2} \rightarrow b_2^*$	0.14(0.16)
$^3B_1$	$3d_{xy} \rightarrow b_2^*$	0.24(0.26)
$^5A_2$	$3d_{x^2-y^2}3d_{z^2} \rightarrow b_1^*b_2^*$	0.52(0.57)
$^5A_1$	$3d_{z^2}3d_{xy} \rightarrow (b_2^*)(b_2^*)$	0.60(0.65)
$^3A_2$	$3d_{z^2}3d_{xy} \rightarrow (b_2^*)^2$	1.11(1.19)
$^5B_1$	$3d_{x^2-y^2}3d_{xy} \rightarrow (b_1^*)(b_2^*)$	1.64(1.77)
$^5B_2$	$3d_{xy}b_2 \rightarrow b_1^*b_2^*$	1.85(1.99)
$^3A_1$	$b_2 \rightarrow (b_2^*)$	1.85(1.93)

[a] Energies obtained with a CASPT2 Co–N_{imido} optimized distance within parenthesis.

3. Results

We shall present two sets of results. First, we present the data obtained with the VDZP basis set where the Cholesky decomposition technique has not been used. Subsequently, we shall illustrate the use of the Cholesky technique and compare results obtained with the VDZP and VTZP basis sets. As we shall see, the two sets of results are very similar and most of the analysis can therefore be done at the VDZP level.

3.1. The Electronic Structure of Co^{III}(diiminato)(NPh).

Here, we describe the different CASPT2 results obtained with the VDZP basis set. The calculations were first performed with a PW91 optimized geometry for the $^1A_1$ state, with a Co−imido distance of 1.653 Å. The 10in10 active space described above was used. CASPT2 calculations were performed for the lowest singlet, triplet, and quintet states in each symmetry. The seven strongly occupied active orbitals for the $^1A_1$ state are shown in Figure 2.

Vertical electronic excitation energies for the lowest triplet and quintet states in each symmetry are presented in Table 1, where we have used the following orbital labels: b_1 and b_1* are the "in-plane" Co−N_{imido} π orbitals and b_2 and b_2* are the "out-of-plane" Co−N_{imido} π orbitals. The CASPT2 calculations yield a closed-shell singlet as the ground state consistent with the observed diamagentism of Co^{III}(nacnac-)(NAd). There are, however, two low-lying triplet excited

Table 2. Vertical CASPT2/VDZP Energies (eV) of the Two Lowest Excited Singlet, Triplet, and Quintet States in Each Symmetry for Co^{III}(diiminato)(NPh) Using an OLYP Optimized Geometry for the Lowest Singlet State

state	configuration	energy[a]
1^1A_1	$(3d_{z^2})^2(3d_{x^2-y^2})^2(3d_{xy})^2(b_1)^2(b_2)^2$	
1^3B_2	$3d_{x^2-y^2} \rightarrow b_2^*$	0.11 (0.14)
1^3B_1	$3d_{xy} \rightarrow b_2^*$	0.15 (0.24)
2^3B_1	$3d_{z^2} \rightarrow b_1^*$	0.52
1^5A_1	$3d_{z^2}3d_{xy} \rightarrow b_1^*b_2^*$	0.59 (0.60)
1^5A_2	$3d_{x^2-y^2}3d_{z^2} \rightarrow b_1^*b_2^*$	0.59 (0.52)
2^3B_2	$3d_{xy} \rightarrow b_1^*$	0.65
1^1B_2	$3d_{x^2-y^2} \rightarrow b_2^*$	0.83
1^1B_1	$3d_{xy} \rightarrow b_2^*$	0.84 (0.84)
1^3A_1	$3d_{z^2}3d_{x^2-y^2} \rightarrow (b_2^*)^2$	0.96 (1.85)[b]
1^3A_2	$3d_{x^2-y^2}3d_{xy} \rightarrow (b_2^*)^2$	1.04 (1.11)
2^1B_2	$0.73(3d_{z^2} \rightarrow b_2^*) - 0.51(3d_{xy} \rightarrow b_1^*)$	1.07
2^3A_2	$3d_{z^2}3d_{x^2-y^2} \rightarrow b_1^*b_2^*$	1.35
2^3A_1	$3d_{z^2}3d_{xy} \rightarrow b_1^*b_2^*$	1.48
2^1B_1	$3d_{x^2-y^2} \rightarrow b_1^*$	1.52
2^5A_1	$3d_{x^2-y^2}3d_{xy} \rightarrow b_1^*b_2^*$	1.54
1^5B_1	$3d_{x^2-y^2}b_2 \rightarrow b_1^*b_2^*$	1.71 (1.64)
1^5B_2	$3d_{x^2-y^2}b_1 \rightarrow (b_1^*)(b_2^*)$	1.89 (1.85)
2^5B_2	$3d_{xy}b_2 \rightarrow b_1^*b_2^*$	1.91
2^5B_1	$3d_{xy}b_1 \rightarrow b_1^*b_2^*$	2.00
2^5A_2	$3d_{z^2}3d_{xy}b_1 \rightarrow b_1^*(b_2^*)^2$	2.56
1^1A_2	$b_1 \rightarrow b_2^*$	3.14
2^1A_1	$0.61(b_1 \rightarrow b_1^*) + 0.56(b_2 \rightarrow b_2^*)$	3.47
2^1A_2	$b_2 \rightarrow b_1^*$	3.90

[a] Values within parentheses are obtained using one root only. [b] Convergence to a different root (cf. Table 1).

states: $^3B_1$ at 0.24 eV and $^3B_2$ at 0.14 eV. A reoptimization of the geometries for these states may well change the relative ordering of the states. Moreover, the energy differences are so small that they are within the error limits of the CASPT2 method.

As previously discussed,[24] the electronic structure of the $^1A_1$ state is of unusual interest. There is a double bond between Co and the imido nitrogen consisting of two π-bonding orbitals, but there is no net Co$-$N$_{\text{imido}}$ σ bond, which is quite extraordinary. As may be seen from the natural orbital occupation numbers in Figure 2, the wave function is somewhat multiconfigurational, with occupation numbers for the antibonding orbitals of 0.13 and 0.22, respectively. The reason for the absence of a σ bond is the double occupancy of the $3d_{z^2}$ orbital, which is energetically more favorable than using it to form a σ bond. Moving electrons from this orbital to the b_1 or b_2 orbitals would enable the formation of a σ bond but at the expense of the π bonds. However, some of the low-lying excited states (cf. Table 2) have the $3d_{z^2}$ orbital singly occupied, resulting in a partial σ bond and less π bonding (e.g., the state 2^3B_1).

In order to study the energetics as a function of the Co$-$N$_{\text{imido}}$ distance, we reoptimized the latter at the CASPT2 level of theory. This led only to a modest change for the ground state: 1.632 Å (CASPT2) instead of 1.653 Å (PW91). The new excitation energies are given within parentheses in Table 1. They are slightly higher than those obtained with the DFT optimized bond distance. The calculations were then performed at the PW91 optimized geometry for the $^3B_2$ state, which has a Co$-$N$_{\text{imido}}$ bond distance of 1.70 Å. This led to the surprising result that the triplet energy, for which the geometry had been optimized with DFT, is higher than the triplet energy calculated at the singlet geometry. The closed-shell $^1A_1$ state is still the lowest state, and the excitation energies have not changed much. In order to study this problem in more detail, a set of calculations were performed where only the Co$-$N$_{\text{imido}}$ distance was varied, keeping all of the other geometry parameters at the singlet geometry. The results are shown in Figure 3 for the three states of lowest energy.

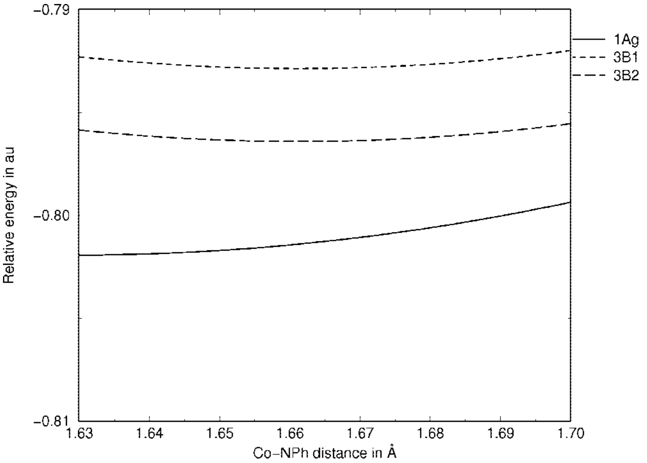


Figure 3. Energy of the three lowest states in the Co^{III}(diiminato)(NPh) complex as a function of the Co$-$N$_{\text{imido}}$ distance.

At the CASPT2/VDZP level, the ground state has a minimum at R(Co$-$N) = 1.632 Å; the next state, $^3B_1$, at 1.661 Å, is considerably shorter than the 1.70 Å of the DFT triplet geometry. The third state, $^3B_2$, has a minimum at the same distance. The three curves are almost parallel, indicating that the excitation energies are not very dependent on the geometry. Of course, this conclusion could change if other geometry parameters were also varied.

CASPT2 calculations were also performed using an OLYP optimized geometry, which is very similar to the one used above and has the same Co$-$N$_{\text{imido}}$ distance (1.653 Å). The results were very similar to those discussed above with one striking difference: the CASPT2 energy of 1.85 eV for the $^3A_1$ state with the PW91 ground-state geometry is now reduced to 0.99 eV (with the OLYP ground-state geometry). The reason for this is convergence to a different root in the CASSCF calculation, which is not surprising in view of the many near-degeneracies involved in this system. Depending on the starting geometry, two different calculations may therefore converge to different roots if they are close in energy. To explore the electronic structure in more detail, we therefore decided to extend the calculations to two roots in each symmetry. In total, 24 electronic states were computed (singlets, triplets, and quintets in each symmetry). The resulting energies are presented in Table 2.

The energies for the first root in each symmetry and spin are very similar to the energies presented in Table 1 with one exception. Comparing the two tables, we see that in one case the single root calculations have converged to the second state instead of the first. This can of course happen, especially when roots are as closely spaced as they are here. The other excitation energies are, however, very similar, and the wave

Table 3. Total CASPT2/VDZP Energies with and without the Cholesky Technique[a]

state	CASPT2	CD-CASPT2	energy difference
1^1A_1	−2364.79544388	−2364.79295568	−0.00248820
2^1A_1	−2364.67390089	−2364.67141410	−0.00248679
1^1B_1	−2364.77033932	−2364.76785555	−0.00248377
2^1B_1	−2364.74534751	−2364.74287943	−0.00246808
1^1B_2	−2364.77072043	−2364.76830977	−0.00241066
2^1B_2	−2364.76208911	−2364.75960783	−0.00248128
1^1A_2	−2364.68597422	−2364.68331489	−0.00265933
2^1A_2	−2364.65783067	−2364.69289360	0.03506293
1^3A_1	−2364.76595240	−2364.76355308	−0.00239932
2^3A_1	−2364.74697718	−2364.74449320	−0.00248398
1^3B_1	−2364.79568110	−2364.79317394	−0.00250716
2^3B_1	−2364.78222016	−2364.77976550	−0.00245466
1^3B_2	−2364.79732094	−2364.79489954	−0.00242140
2^3B_2	−2364.77725821	−2364.77474030	−0.00251791
1^3A_2	−2364.76313971	−2364.76067077	−0.00246894
2^3A_2	−2364.75177134	−2364.74936282	−0.00240852
1^5A_1	−2364.77946626	−2364.77696359	−0.00250267
2^5A_1	−2364.74481204	−2364.74236741	−0.00244463
1^5B_1	−2364.73843766	−2364.73598410	−0.00245356
2^5B_1	−2364.72771500	−2364.72518246	−0.00253254
1^5B_2	−2364.73200864	−2364.72953416	−0.00247448
2^5B_2	−2364.73100948	−2364.72850268	−0.00250680
1^5A_2	−2364.77947072	−2364.77704933	−0.00242139
2^5A_2	−2364.70713998	−2364.70466298	−0.00247700

[a] A threshold of 10^{-4} was used for the Cholesky decomposition.

Table 4. Timing information for the Cholesky calculations[a]

	VDZP basis		
	conventional	Cholesky	VTZP basis
integrals	150	56	217
CASSCF/it[b]	4.25	0.41	3.88
CASPT2	383 (24)	163 (20)	898 (232)

[a] Wall times are given in minutes (CPU time within parentheses for the CASPT2 calculations). [b] Wall time per CASSCF iteration.

Table 5. A Comparison of the CASPT2 Relative Energies (eV) with the VDZP and VTZP Basis Sets

state	VTZP	VDZP	state	VTZP	VDZP
1^1A_1	0.00	0.00	2^3A_2	1.39	1.35
1^3B_1	0.02	0.15	2^1B_1	1.55	1.52
1^3B_2	0.18	0.11	2^5A_1	1.57	1.54
2^3B_1	0.59	0.52	1^5B_1	1.70	1.71
2^3B_2	0.63	0.65	1^5B_2	1.98	1.89
1^5A_1	0.63	0.59	2^3A_1	2.03	1.48[a]
1^5A_2	0.64	0.59	2^5B_1	2.40	2.00[a]
1^1B_2	0.86	0.83	2^5B_2	2.45	1.91[a]
1^1B_1	0.87	0.84	2^5A_2	2.59	2.56
1^3A_1	0.90	0.96	2^1A_2	2.82	3.90[a]
2^1B_2	1.04	1.07	1^1A_2	3.13	3.14
1^3A_2	1.05	1.04	2^1A_1	3.46	3.47

[a] Convergence to different roots.

Table 6. Relative Electronic Energies E_{rel}, Thermodynamic Energies (U_{rel}),[b] or Enthalpies (H_{rel})[c] and Relative Gibbs Free Energies G_{rel} Based on OLYP/STO-TZP Geometries and Harmonic Frequencies and the Ideal Gas Approximation at 298.15 K[a]

state	E_{rel}	U_{rel}/H_{rel}	G_{rel}
1^1A_1	0.00	0.00	0.00
1^3B_2	5.38(0.23)	4.96(0.21)	4.43(0.19)
1^3B_1	10.67(0.46)	10.17(0.44)	9.29(0.40)
2^3B_1	13.90(0.60)	11.78(0.51)	12.84(0.56)

[a] The energies shown are in kcal/mol (eV). [b] The thermodynamic energy U is the sum of the electronic energy E, the zero-point energy, and translational, vibrational, and rotational energies at 298.15 K. [c] Trends in U_{rel} and H_{rel} are identical since the work term RT in the latter cancels out.

functions have the same orbital occupancy. We can therefore limit ourselves to the two-root calculations in our further discussion.

Two triplet states have very low energies: $^3B_1$ and $^3B_2$. Their energies are so close to zero that we cannot conclusively determine the real ground state in this system. The accuracy of these calculations (or any other calculation) is certainly not better than 0.1 eV. Our results thus indicate that three electronic states are very close in energy, and they are all good candidates for the ground state. The density of states is very high: there are 20 electronic states below 2.0 eV. Several of these are doubly excited with respect to the $^1A_1$ ground state. The reason for this high density of states is that there are three doubly occupied 3d orbitals, which are very close in energy in the ground state: $3d_{z^2}$, $3d_{x^2-y^2}$, and $3d_{xy}$.

3.2. CD-CASPT2 Calculations. In this section, we shall study the effect of the basis set on the results obtained above. To perform calculations with a VTZP basis set (869 ANO-RCC functions), we needed to invoke the Cholesky decomposition technique. We started by calibrating this methodology against full integral calculations with the VDZP basis set used above. The results of this comparison are shown in Table 3, which gives the total CASPT2 energies obtained with and without CD. A threshold of 10^{-4} was used in decomposing the ERI matrix. The energy difference between the two sets of calculations is almost constant and on the order of 2×10^{-3} E_h (about 0.07 eV), which is an order of magnitude larger than the threshold used. More importantly, the relative energies of the states are affected by less than 0.01 eV. [There is one exception: the 2^1A_2 state, where the CD calculation converged to a different electronic state.] Calculations with a threshold of 10^{-8} gave absolute energies with an accuracy of about 10^{-6} au and relative energies identical to those obtained with full integrals. We therefore conclude that the CD calculations are accurate. On the basis of our earlier experience, we would expect the accuracy to be even better with the larger basis set.[22]

Table 4 presents the wall-clock times obtained on an AMD Opteron 148 (2.2 GHz) PC for one single-root calculation ($^1A_1$). With the VDZP basis set, the generation of the Cholesky vectors is 3 times faster than the full integral calculation. The increase in speed is much larger for the VTZP basis set, where the calculation of the Cholesky vectors takes 217 min. The full integral calculation could not be performed in this case, but it may be estimated to be around 4500 min from the timing obtained for the VDZP basis set. In other words, the CD method results in a speedup by about a factor of 20. The overall CASSCF/VDZP calculations are a factor of 10 faster, a speedup that should also increase for the larger basis set. For the VTZP basis set, each CASSCF iteration takes 3.88 min, underscoring the effectiveness of the CD technique. The speedup of the CASPT2 calculations is less impressive. Not much is gained in CPU time, and I/O times overwhelmingly dominate the calculations. The reason for this was explained in the

Table 7. Comparison of Vertical Excitation Energies (eV) with CASPT2 and DFT

		CASPT2		DFT(STO-TZP)						
	electronic configuration	VDZP	VTZP	OLYP	OPBE	BLYP	PW91	BP86	B3LYP	B3LYP*
1^1A_1	$(3d_{z^2})^2(3d_{x^2-y^2})^2(3d_{xy})^2(b_1{}^*)^0(b_2{}^*)^0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1^3B_2	$(3d_{z^2})^2(3d_{x^2-y^2})^1(3d_{xy})^2(b_1{}^*)^0(b_2{}^*)^1$	0.11	0.18	0.31	0.19	0.52	0.44	0.43	−0.09	0.07
1^3B_1	$(3d_{z^2})^2(3d_{x^2-y^2})^2(3d_{xy})^1(b_1{}^*)^0(b_2{}^*)^1$	0.15	0.02	0.60	0.55	0.70	0.67	0.66	0.06	0.23
2^3B_1	$(3d_{z^2})^2(3d_{x^2-y^2})^1(3d_{xy})^2(b_1{}^*)^1(b_2{}^*)^0$	0.52	0.59	0.88	0.82	1.08	1.04	1.03	0.24	0.51
2^3B_2	$(3d_{z^2})^2(3d_{x^2-y^2})^2(3d_{xy})^1(b_1{}^*)^1(b_2{}^*)^0$	0.65	0.63	1.35	1.31	1.50	1.46	1.45	0.54	0.83
1^5A_2	$(3d_{z^2})^1(3d_{x^2-y^2})^1(3d_{xy})^2(b_1{}^*)^1(b_2{}^*)^1$	0.59	0.64	1.01	0.84	1.47	1.35	1.34	0.33	0.63
1^5A_1	$(3d_{z^2})^2(3d_{x^2-y^2})^1(3d_{xy})^1(b_1{}^*)^1(b_2{}^*)^1$	0.59	0.63	1.19	1.06	1.63	1.53	1.51	0.47	0.77
max deviation[a]				0.72	0.68	0.87	0.90	0.88	0.35	0.21
rms deviation[a]				0.44	0.38	0.68	0.62	0.61	0.22	0.13

[a] Deviations from the CASPT2(VTZP) results.

Methodology section, and we expect that future versions of the MOLCAS program system will result in substantially improved timings. Nevertheless, we have demonstrated that a CASSCF/CASPT2 calculation with 869 ANO-RCC basis functions can be performed on a normal PC in less than 20 h of wall-clock time.

The results for the VTZP calculation are presented in Table 5 where they are compared with the VDZP results described above. Generally, the results are very similar. In a few cases the VTZP calculations of the second root converged to a different state, but we shall not study these cases here. They only mean that there are more low-lying states of the symmetry in question, and to obtain all of them correctly, one would have to increase the number of states included in the calculations. In most other cases, the energy differences are less than 0.1 eV with one exception. The 1^3B_1 state is now the second state and is virtually degenerate with the 1^1A_1 state. Remembering that the accuracy of the CASPT2 method is in general no better than about 0.2 eV, these results make it impossible to conclusively determine the ground state. All that can be concluded is that there are three candidates: 1^1A_1, 1^3B_1, and 1^3B_2.

3.3. Free Energy Differences. The question of definitively identifying the ground state cannot be addressed without including the zero-point energy and entropic effects. We have therefore computed these quantities for four of the lowest-energy states, using OLYP/STO-TZP geometries and harmonic frequencies and the ideal gas approximation at 298.15 K. The results are presented in Table 6. We note that these adiabatic relative energies are different from the vertical ones presented above. The changes, however, are small. Not surprisingly, geometry optimization lowers the energy of the triplet states, relative to the ground state, the margins being −0.04 eV for 1^3B_2, −0.06 eV for 1^3B_1, and −0.04 eV for 2^3B_1. Overall, according to Table 6, we still have at least two and possibly three near-degenerate lowest-energy states, and we cannot tell, on the basis of calculations alone, which should be the actual ground state.

4. Comparing CASPT2 and DFT

The CASPT2 results, which we estimate to be accurate to about ±0.2 eV, provide valuable calibration for DFT calculations of the energies of several of the low-lying spin states of Co^{III}(diiminato)(NPh). Seven low-lying states were calculated with seven popular exchange-correlation functionals at the OLYP, $S = 0$, ground-state geometry. Table 7 presents a comparison of the DFT/TZP and CASPT2 energetics. According to Table 7, the B3LYP* functional, which has a reduced amount (15%) of Hartree–Fock exchange relative to B3LYP (20%), appears to be in the best agreement with CASPT2. The more widely used B3LYP functional performs slightly worse, predicting (albeit by a margin of only 0.09 eV) a triplet $^3B_2$ ground state,[51] which is inconsistent with the experimentally observed diamagnetism of Co^{III}(nacnac)(NAd). In contrast, the classical pure functionals BLYP, PW91, and BP86 appear to exaggerate the instability of the higher-multiplicity states. The newer OPTX-based pure functionals[25] OLYP and OPBE are somewhat better in this respect. Between OLYP and OPBE, the energetics provided by OPBE appears to be in slightly better agreement with CASPT2 than those obtained with OLYP, as indicated by the slightly smaller root-mean-square (rms) error. The maximum deviations are, however, almost as large for these functionals as they are for the other pure functionals. In summary, among the functionals examined, B3LYP and in particular B3LYP* seem to be the only that are acceptable, as far as the spin-state energetics of Co^{III}-(diiminato)(NPh) are concerned. It will be interesting to see how well this conclusion might generalize as additional complexes are examined in similar studies. That said, there are many examples in the literature where B3LYP (or B3LYP*) has performed no better than and even distinctly worse than pure functionals.[29–40,52] The development of new, broadly applicable functionals is therefore very much an ongoing process.

5. Conclusions

This work, in our opinion, has contributed on three different fronts.

First, we have shown how CD of the two-electron integral matrix can be used to drastically simplify electronic structure calculations of large molecules and with accurate basis sets and wave-function-based methods. The CD method has been used in all stages of the calculations from SCF and CASSCF to CASPT2. The current implementation of this technique is not yet optimal but still allows calculations with nearly (and almost certainly somewhat over) 1000 basis functions. In the future, we expect to be able to use similarly accurate basis sets in ab initio studies of considerably larger molecules. Further improvement of the CD-CASPT2 part of our code[50] is needed for achieving this, but the technology is already at hand. The 43-atom molecule studied here was treated with an ANO-RCC-VTZP basis set. With a Cholesky threshold of 10^{-4}, we can assume an accuracy better than 0.01 eV in the relative energies for 24 electronic states (this accuracy refers to the error introduced by the finite Cholesky function threshold). The generation of the Cholesky vectors with this threshold and the VTZP basis set is estimated to be about 20 times faster than the full generation of the ERI matrix. To this should be added the savings in disk storage requirement, which is reduced by several orders of magnitude. The large savings in computer time and resources combined with the maintained accuracy strongly suggest that the CD-CASPT2 approach will be the standard in future applications.

Second, CASPT2 calculations of the spin-state energetics of a Co^{III}(diiminato)(NPh) complex have provided valuable calibration of analogous DFT calculations. Surprisingly, such calibrations of DFT (vis-a-vis the specific issue of transition metal spin-state energetics) are quite rare. In this study, we have calculated seven low-lying spin states of Co^{III}(diiminato)(NPh) with seven common exchange-correlation functionals and compared the results with CASPT2. The B3LYP* functional, containing a reduced amount (15%) of exchange relative to the more widely used B3LYP functional, appears to be the best. Among the pure functionals examined, the newer OPTX-based functionals OPBE and OLYP appear to perform slightly better than older, classic functionals such as PW91, BLYP, and BP86 but still give unacceptably large rms error for the computed excitation energies.

Last, this study has also deepened our growing understanding of bonding in low-coordinate imido complexes. That all known Co^{III}−imido complexes exhibit diamagnetic ground states certainly appears to be a coincidence in light of the results obtained in this study. Thus, at least for Co^{III}−diiminato−imido complexes, our calculations predict multiple paramagnetic excited states at very low energies (perhaps as low as a couple of tenths of an electrovolt) above the ground state. This is an important facet of the electronic structure of these complexes that has not yet manifested itself in experimental studies.[23,28] However, as already mentioned, one or more thermally accessible paramagnetic excited states have been implicated for a hydrotrispyrazolylborate-supported Co^{III}−imido complex.[26–28]

Acknowledgment. This work has been supported by the Swedish Natural Science Research Council (VR) and by the Research Council of Norway. A.G. thanks Dr. Jeanet Conradie and Mr. Espen Tangen for assistance with the DFT calculations.

References

(1) Although translated tongue-in-cheek as a success that has run out of steam, this French expression is not a slur. Instead, it refers to a significant success that is appreciated by connoisseurs but lacks much of a popular following.

(2) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. In *Advances in Chemical Physics: New Methods in Computational Quantum Mechanics*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons: New York, 1996; Vol. XCIII, pp 219–332.

(3) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094.

(4) Ghosh, A.; Taylor, P. R. *Curr. Opin. Chem. Biol.* **2003**, *91*, 113–124.

(5) Roos, B. O.; Borin, A. C.; Gagliardi, L. *Angew. Chem., Int. Ed.* **2006**, *46*, 1469–1472.

(6) Noodleman, L.; Lovell, T.; Han, W.-G.; Li, J.; Himo, F. *Chem. Rev.* **2004**, *104*, 459–508.

(7) Siegbahn, P. E. M.; Borowski, T. *Acc. Chem. Res.* **2006**, *39*, 729–738.

(8) Ghosh, A.; Steene, E. J. *Biol. Inorg. Chem.* **2001**, *6*, 739–752.

(9) Ghosh, A. *Acc. Chem. Res.* **2005**, *38*, 943–954.

(10) Roos, B. O. 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.

(11) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. J. *Phys. Chem.* **1990**, *94*, 5483–5488.

(12) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(13) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *7*, 683–705.

(14) Røeggen, I. R.; Wisløff-Nielsen, E. *Chem. Phys. Lett.* **1986**, *132*, 154–160.

(15) Koch, H.; Sánchez de Merás, A.; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481–9484.

(16) Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, Art. no. 194106.

(17) Pedersen, T. B.; Sánchez de Merás, A. M. J.; Koch, H. *J. Chem. Phys.* **2004**, *120*, 8887–8897.

(18) Pedersen, T. B.; Koch, H.; Boman, L.; Sánchez de Merás, A. M. *J. Chem. Phys. Lett.* **2004**, *393*, 319–326.

(19) García Cuesta, I.; Pedersen, T. B.; Koch, H.; Sánchez de Merás, A. *ChemPhysChem* **2006**, *7*, 2503–2507.

(20) Fernández, B.; Pedersen, T. B.; Sánchez de Merás, A.; Koch, H. *Chem. Phys. Lett.* **2007**, *441*, 332–335.

(21) Öhrn, A.; Aquilante, F. *Phys. Chem. Chem. Phys.* **2007**, *9*, 470–480.

(22) Aquilante, F.; Pedersen, T. B.; Roos, B. O.; Sánchez de Merás, A.; Koch, H. *J. Chem. Phys.*, **2007**, submitted.

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

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

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

(26) Shay, D. T.; Yap, G. P. A.; Zakharov, L. N.; Rheingold, A. L.; Theopold, K. H. *Angew. Chem., Int. Ed.* **2005**, *44*, 1508–1510.

(27) Shay, D. T.; Yap, G. P. A.; Zakharov, L. N.; Rheingold, A. L.; Theopold, K. H. *Angew. Chem., Int. Ed.*, **2006**, *45*, 7870–7870. Corrigendum.

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

(29) Reiher, M.; Salomon, O.; Hess, B. *Theor. Chem. Acc.* **2001**, *107*, 48–51.

(30) Swart, M.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2004**, *108*, 5479–5483.

(31) Swart, M.; Ehlers, A. W.; Lammertsma, K. *Mol. Phys.* **2004**, *102*, 2467–2474.

(32) Deeth, R. J.; Fey, N. *J. Comput. Chem.* **2004**, *25*, 1840–1848.

(33) Groenhof, A. R.; Swart, M.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2005**, *109*, 3411–3417.

(34) Daku, L. M. L.; Vargas, A.; Hauser, A.; Fouqueau, A.; Casida, M. E. *ChemPhysChem* **2005**, *6*, 1393–1410.

(35) Ganzenmuller, G.; Berkaine, N.; Fouqueau, A.; Casida, M. E.; Reiher, M. J. *J. Chem. Phys.* **2005**, *0122*, Art. No. 234321.

(36) De Angelis, F.; Jin, N.; Car, R.; Groves, J. T. *Inorg. Chem.* **2006**, *45*, 4268–4276.

(37) Vargas, A.; Zerara, M.; Krausz, E.; Hauser, A.; Daku, L. M. L. *J. Chem. Theory Comput.* **2006**, *2*, 1342–1359.

(38) Rong, C. Y.; Lian, S. X.; Yin, D. L.; Shen, B.; Zhong, A. G.; Bartolotti, L.; Liu, S. B. *J. Chem. Phys.* **2006**, *125*, Art. No. 174102.

(39) Strickland, N.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841–852.

(40) Conradie, J.; Ghosh, A. *J. Phys. Chem. B* **2007**, *111*, 12621–12624.

(41) Mehn, M. P.; Peters, J. C. *J. Inorg. Biochem.* **2006**, *100*, 634–643.

(42) O'Neal, D. W.; Simons, J. *Int. J. Quantum Chem.* **1989**, *36*, 673–688.

(43) Wilson, S. *Comput. Phys. Commun.* **1990**, *58*, 71–81.

(44) García Cuesta, I.; Pedersen, T. B.; Koch, H.; Sánchez de Merás, A. M. *J. Chem. Phys. Lett.* **2004**, *390*, 170–175.

(45) Velde, G. T.; Baerends, E. J.; Guerra, C. F.; Gisbergen, S. J. A. V.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967. The ADF program system was obtained from Scientific Computing and Modeling, Amsterdam (http://www.scm.com).

(46) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851.

(47) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.

(48) Aquilante, F.; Pedersen, T. B.; Koch, H.; Sánchez de Merás, A. *J. Chem. Phys.* **2006**, *125*, Art. no. 174101.

(49) Andersson, K.; Roos, B. O. *Chem. Phys. Lett.* **1992**, *191*, 507.

(50) 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. *Comp. Mater. Sci.* **2003**, *28*, 222–239.

(51) It should be mentioned that, if the same DFT/B3LYP calculations are performed using the ANO-RCC-VDZP basis set, including scalar relativistic effects, the 1^3B_2 state will appear 0.10 eV below the $^1A_1$ ground state. The energy difference for the 1^3B_1 state is +0.03 eV.

(52) Sorkin, A.; Iron, M. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 307–315.

CT700263H

JCTC Journal of Chemical Theory and Computation

Electronic Structures of $AlGaN_2$ Nanotubes and AlN-GaN Nanotube Superlattice

Hui Pan,*,† Yuan Ping Feng,*,† and Jianyi Lin‡

*Department of Physics, National University of Singapore, 2 Science Drive 2,
Singapore 117542, and Institute of Chemical and Engineering Sciences, 1 Pesek Road,
Jurong Island, Singapore 627833*

Received November 14, 2007

Abstract: The electronic properties of single-wall $AlGaN_2$ nanotubes were investigated using first-principles calculations and generalized gradient approximation. All $AlGaN_2$ nanotubes considered are semiconductors, but their band structures depend on their chirality and size due to curvature effect and symmetry. The zigzag $AlGaN_2$ nanotubes are direct band gap semiconductors, while armchair $AlGaN_2$ nanotubes are indirect band gap semiconductors. The calculations on the electronic properties of AlN-GaN nanotubes superlattice show that the band gap engineering can be realized by changing the composition of the AlN-GaN nanotubes superlattice.

Introduction

Nanotubes have attracted extensive attention for their intriguing and potentially useful structural, electrical, and mechanical properties since the discovery of carbon nanotube (CNT). Theoretically, a number of nanotubes, such as GaN,[1] BN,[2] WC,[3] BC_2N,[4] SiC,[5] and AlN[6] nanotubes, have been predicted. Experimentally, a variety of nanotubes, such as BN,[7–9] $B_xC_yN_z$,[10,11] and AlN,[12] have been successfully synthesized by various methods, such as pulsed laser deposition, chemical vapor deposition, and wet chemistry. Recently, M. Remskar classified the inorganic nanotube (NTs) into six groups, including the following: oxide NTs, transition-metal chalcogenide NTs, transition-metal halogenous NTs, mixed-phase and metal-doped NTs, carbon-, boron-, and silicon-based NTs, and metal NTs.[13] For example, R. Tenne et al. first reported the transition-metal chalcogenide NTs, WS_2, and MoS_2, in 1992[14] and 1995,[15] respectively, and the transition-metal halogenous NTs, $NiCl_2$, in 1998.[16] They also studied the mechanical property and Raman scattering of WS_2 NTs.[17,18] It is well-known that the electronic properties of nanotubes depend on the size (radius) and chirality of the nanotubes. For single-walled CNTs (SWCNTs), the band gap of semiconducting SWCNT is inversely proportional to its diameter.[19] As for BC_2N nanotube, recent calculations indicated that both its electronic and optical properties were size and chirality dependent.[4,20]

III−V compound semiconductors are important materials in device application. Theoretical calculations indicated that the band gap of AlN and GaN single-wall nanotubes can be controlled by varying the size and chirality,[1,7] suggesting the applicability to full color flat panel displays. Experimentally, a bulk ternary semiconductor ($Al_xGa_{1-x}N$) has been widely studied for its application in devices, such as quantum well devices.[21] Gudiksen et al. reported the compositionally modulated superlattice nanowires consisting of 2−21 layers of GaAs and GaP for nanoscale photonics and electronics.[22] The superlattices are created within the nanowires by repeated modulation of the vapor-phase semiconductor reactants during growth of the wires. Multielement nanotubes can be expected to provide more tenability to their physical properties and to meet requirements of various applications. To the best of our knowledge, theoretically, the ternary nanotube, including Al, Ga, and N, has not been studied. In this article, we investigate the chirality and size dependence of electronic properties of armchair and zigzag $AlGaN_2$ nanotubes. We

* Corresponding author e-mail: phyph@nus.edu.sg (H.P.); phyfyp@nus.edu.sg (Y.P.F.).

† National University of Singapore.

‡ Institute of Chemical and Engineering Sciences.

10.1021/ct7003116 CCC: $40.75
Published on Web 04/15/2008

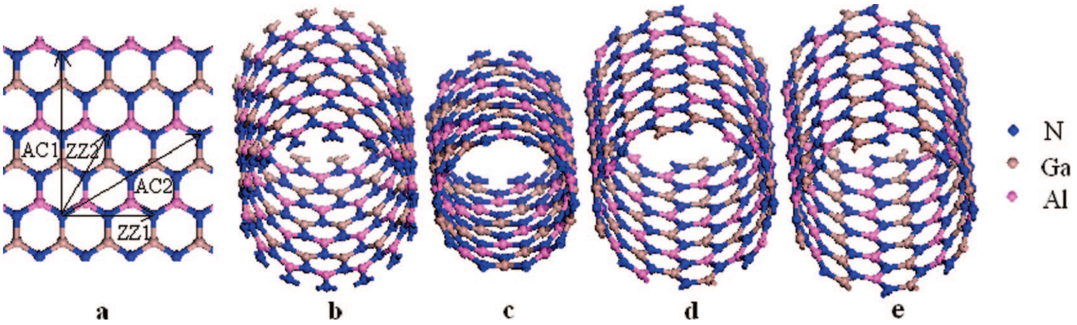


Figure 1. Atomic configurations of (a) the most stable $AlGaN_2$ sheet, (b) ZZ1 (14,0), (c) ZZ2 (0,5), (d) AC1 (8,8), and (e) AC2 (4,4) $AlGaN_2$ nanotubes. The Ga, Al, and N atoms are indicated by brown, pink, and blue spheres. The wrapping vectors of the four types of nanotubes are shown in (a).

also study possible band gap engineering by varying the composition of an AlN-GaN nanotube superlattice.

Calculation Details

We carried out first-principles calculation based on the density functional theory (DFT)[23] and the generalized gradient approximation (GGA).[24] The plane-wave based pseudopotential method and the CASTEP code were used in the study.[25] The ionic potentials are described by the ultrasoft nonlocal pseudopotential proposed by Vanderbilt.[26] The Monkhorst and Pack scheme of k point sampling was used for integration over the first Brillouin zone.[27] The Kohn−Sham energy functional is directly minimized using the conjugate-gradient method.[28] The convergence test indicated that an energy cutoff of 350 eV was sufficient for the calculations.

Compared to carbon nanotubes, there can be more than one type of zigzag or armchair $AlGaN_2$ nanotubes, depending on how an $AlGaN_2$ sheet is rolled up (Figure 1a). In this study, we considered two types of zigzag nanotubes: ZZ1 $(n, 0)$ with $n = 5-16$ and ZZ2 $(0, n)$ with $n = 3-8$, and two types of armchair nanotubes: AC1 (n, n) with $n = 3-11$ and AC2 (m, m) with $m = 2-5$, as shown in Figures 1b−e. In addition, AlN-GaN (6, 0) nanotube superlattices were studied. As the cell dimension in the direction of tube axis is different for different tube chirality, the k points used in the calculations are adjusted accordingly so that its density in the reciprocal space remains more or less the same, and the number of k points used in the calculations are 12 for ZZ1, 6 for ZZ2, 14 for AC1, and 6 for AC2 types of $AlGaN_2$ nanotubes, respectively. The k points used in the calculations of the AlN-GaN nanotube superlattices are 4. Good convergence was obtained with these parameters, and the total energy was converged to 2.0×10^{-5} eV/atom. A large supercell dimension with a wall−wall distance of 10 Å in the plane perpendicular to the tube axis was used to avoid interaction between the nanotube and its images in neighboring cells. The unit is periodic in the direction of the tube. The geometrically optimized nanotubes were used for band structure and optical property calculations.

As an indication of stability, the binding energy is estimated from the formula

$$E_b = |E_{\mathrm{tube}} - n\mu_{\mathrm{Al}} - n\mu_{\mathrm{Ga}} - 2n\mu_{\mathrm{N}}| \qquad (1)$$

Table 1. Bond Lengths in Different Nanotubes after Geometry Optimization

nanotube	Al−N (Å)	Ga−N (Å)
AlN	1.81	
GaN		1.86
$AlGaN_2$	1.76	1.84
AlN-GaN	1.80	1.87

where E_{tube} is the energy of the $AlGaN_2$ nanotube. μ_{Al}, μ_{Ga}, and μ_{N} are chemical potentials of Al, Ga, and N atoms, respectively. n is the number of Al (or Ga) atoms in the nanotube.

Results and Discussion

A number of possible structures for planner $AlGaN_2$ were considered. Our total energy calculations indicated that the geometry with Ga and Al atoms separated by N atoms (Figure 1a) is most stable due to the lowest total energy. Other structures, such as the one with Ga atoms bonding to Al atoms, are less stable than that shown in Figure 1a due to the higher energy. The covalent bond lengths in the fully optimized structures are given in Table 1. For AlN and GaN nanotubes, the structure details are in good agreement with those of refs 1 and 6. The Al−N and Ga−N bond lengths in $AlGaN_2$ nanotubes are slightly less than those in AlN and GaN nanotubes.

Figure 2 shows the total energies per $AlGaN_2$ unit of the optimized $AlGaN_2$ nanotubes as a function of the tube diameter. The energy of the corresponding $AlGaN_2$ sheet is also shown for comparison. We can see that the total energies of all four types of $AlGaN_2$ nanotubes converge to that of the $AlGaN_2$ sheet as the diameter of the tubes increases. The energy difference between the tube and sheet decreases from 0.58 to 0.04 eV with the increase of the tube diameter. Furthermore, the total energies per $AlGaN_2$ unit of all four types of $AlGaN_2$ nanotubes with the same size are essentially the same, indicating that the strain energy of an $AlGaN_2$ nanotube, defined as the energy difference between the $AlGaN_2$ nanotube and the $AlGaN_2$ sheet, does not depend on its chirality. Figure 3 shows the binding energies of the optimized $AlGaN_2$ nanotubes as a function of the tube diameter. Similarly, at the same diameter, the binding energies of the four types of the $AlGaN_2$ nanotubes are almost equal, i.e., the binding energy of the $AlGaN_2$ nanotube

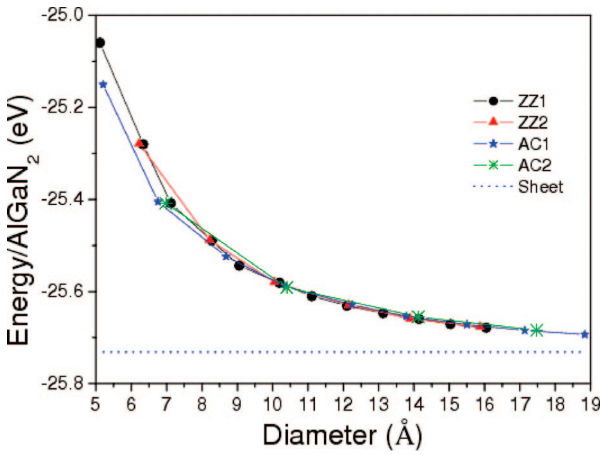


Figure 2. The total energies of $AlGaN_2$ nanotubes as a function of the diameter and a $AlGaN_2$ sheet.

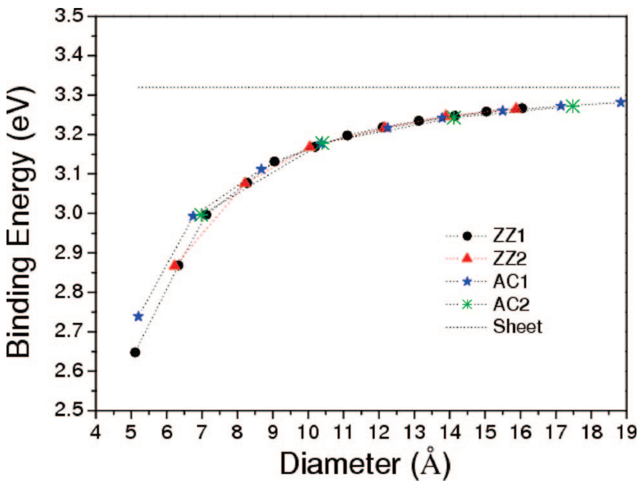


Figure 3. The binding energies of $AlGaN_2$ nanotubes as a function of the diameter and a $AlGaN_2$ sheet.

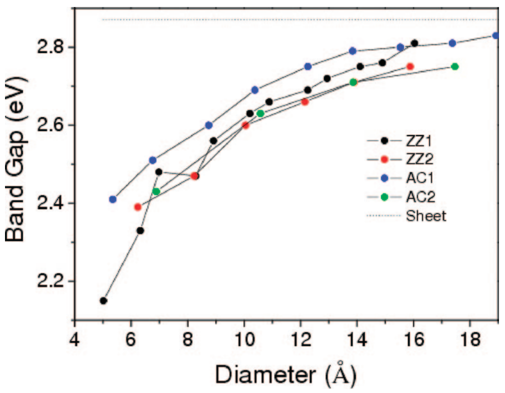


Figure 4. Band gaps of the $AlGaN_2$ nanotubes and the $AlGaN_2$ sheet are shown as functions of their diameters.

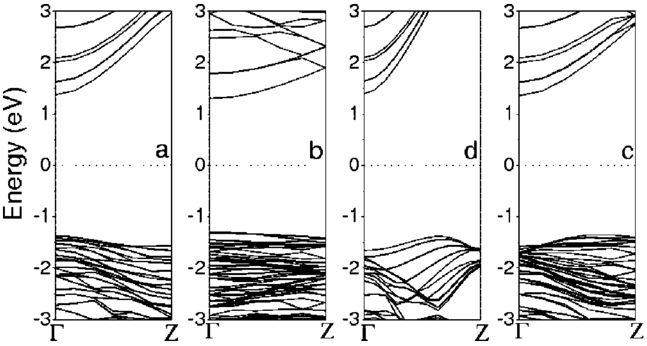


Figure 5. Calculated band structures of (a) ZZ1 (14, 0), (b) ZZ2 (0, 5), (c) AC1 (8, 8), and (d) AC2 (4, 4). The insets show the electron density of the highest valence energy level.

is chirality-independent. However, the binding energy of the $AlGaN_2$ nanotube is size-dependent due to the curvature effect. The binding energy increases with the increase of the diameter, indicating the $AlGaN_2$ nanotubes with larger diameter are more stable than those with smaller diameter. Therefore, from the energy point of view, all four types of $AlGaN_2$ nanotubes may be produced experimentally, although it is easier to grow $AlGaN_2$ nanotubes with larger diameters due to lower strain energy and higher binding energy.

Figure 4 shows the variation of the calculated GGA band gaps of various $AlGaN_2$ nanotubes with the diameter of the tubes. First of all, all $AlGaN_2$ nanotubes considered are semiconductors and the band gap of $AlGaN_2$ nanotube depends on its diameter and chirality. The band gap increases with an increase of diameter and converges to that of the $AlGaN_2$ sheet (2.87 eV) when the diameter of the tube becomes very large. The relatively smaller band gaps for the $AlGaN_2$ nanotubes with smaller diameters can be attributed to the curvature-induced strong hybridization effect. For nanotubes with the same diameter, the AC1 nanotubes have a slightly larger band gap. The band gaps of the $AlGaN_2$ nanotube are generally less than those of AlN and GaN nanotubes.[1,6] And the band gaps of the $AlGaN_2$ nanotubes are less than those of bulk wurtzite $Al_xGa_{1-x}N$ ($0 \leq x \leq 1$) alloys, which have tunable direct band gaps between 3.4 and 6.1 eV, depending on the Al content. Therefore, these $AlGaN_2$ nanotubes can be recognized as important semiconductors for optoelectronic device applications over the visible spectral range.

The representative band structures of the four types of $AlGaN_2$ nanotubes near the Fermi level were demonstrated in Figure 5. All zigzag (ZZ1 and ZZ2) nanotubes are direct band gap semiconductors with the bottom of the conduction energy level and the top of the valence energy level located at the Brillouin zone center (Γ) (Figure 5a,b). On the contrary, all armchair (AC1 and AC2) nanotubes are indirect gap semiconductors, with the bottom of the conduction energy level located at the Γ point but the top of the valence energy level at $\sim{}^{2}/_{3}$ along the ΓZ direction (Figure 4c,d). Analysis of electron densities corresponding to the top valence band of the zigzag $AlGaN_2$ nanotube shows that the top valence band consists of *p* orbitals of the nitrogen atoms next to Al atoms in the direction of the tube axis. These *p* orbitals are normal to the tube surface (Figure 6a). For the armchair $AlGaN_2$ nanotube, the valence top level is attributed to similar *p* orbitals of all nitrogen atoms (Figure 6b). These observations indicate that the electronic properties of the $AlGaN_2$ nanotubes are chirality-dependent. And the valence top levels in $AlGaN_2$ nanotubes with different chirality are attributed to the *p* orbitals from different atoms due to the difference in symmetry.

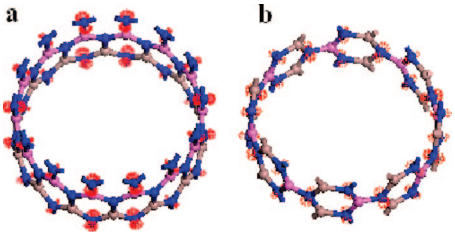


Figure 6. The electron density of the valence top level of (a) ZZ1 (14, 0) and (b) AC1 (8, 8).

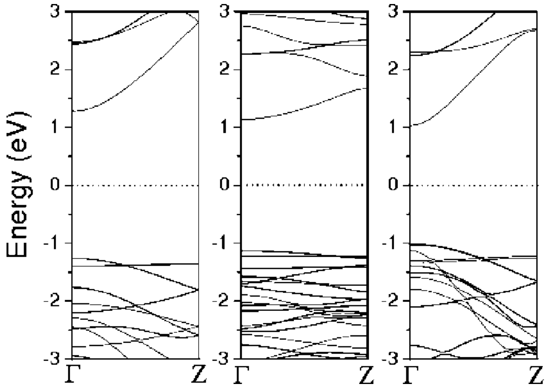


Figure 7. Calculated band structures of (a) AlN (6, 0) nanotube, AlN-GaN (6, 0) nanotube superlattice, and (c) GaN (6, 0) nanotube.

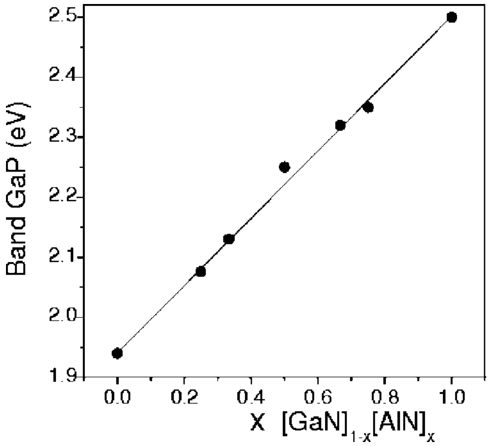


Figure 8. Band gap of [AlN (6, 0)]$_x$[GaN (6, 0)]$_{1-x}$ nanotube superlattice as a function of the x.

The AlN-GaN nanotube superlattices consists of alternating AlN (6, 0) and GaN (6, 0) segments of different lengths, i.e. [AlN (6, 0)]$_x$[GaN (6, 0)]$_{1-x}$. Restuls of our calculations indicated that Al−N and Ga−N bond lengths in the superlattice are very close to those in separate AlN and GaN nanotubes (Table 1). The band structures of AlN (6, 0) nanotube, AlN-GaN nanotube superlattice with $x = 0.5$ and a GaN (6, 0) nanotube, illustrated that they are direct semiconductors (Figure 7). The band gap of the [AlN (6, 0)]$_{0.5}$[GaN (6, 0)]$_{0.5}$ nanotube (2.25 eV) is slightly smaller than that of AlGaN$_2$ (6, 0) nanotube (2.33 eV), which is about the average of the band gaps of the AlN (6, 0) nanotube ($E_{\text{AlN}} = 2.50$ eV) and the GaN (6, 0) nanotube ($E_{\text{GaN}} = 1.94$ eV). The p orbitals normal to the tube surface of the nitrogen atoms at the interface of the junction contribute to the top valence band by the analysis of electron densities of the top valence band of AlN-GaN nanotube superlattice. A number of [AlN (6, 0)]$_x$[GaN (6, 0)]$_{1-x}$ nanotube superlattice with different x were investigated. Figure 8 shows the change of the band gap of the AlN-GaN superlattices (E_s) with x. The linear dependence of E_s on the x implies that E_s for a [AlN (6, 0)]$_x$[GaN (6, 0)]$_{1-x}$ nanotube superlattice with an arbitrary x can be estimated based on the simple linear interpolation, $E_s = xE_{\text{GaN}} + (1\text{-}x)E_{\text{AlN}}$.

Conclusions

In summary, first-principles calculations on the electronic properties of single-wall AlGaN$_2$ nanotubes indicated that the electronic properties of the AlGaN$_2$ nanotubes are size and chirality dependent due to the curvature effect and symmetry. The band gap of the AlGaN$_2$ nanotube increases with increasing size and converges to that of the planner AlGaN$_2$. Our calculations also predicted that the band gap of [AlN (6, 0)]$_x$[GaN (6, 0)]$_{1-x}$ nanotube superlattice can be engineered by changing the composition. Although the well-known fact that DFT/GGA underestimates the band gap of semiconductors, the dependence of the electronic properties of the nanotubes on their size and chirality are valid. The theoretical results should be confirmed experimentally.

References

(1) Lee, S. M.; Lee, Y. H.; Hwang, Y. G.; Elsner, J.; Porezag, D.; Frauenheim, T. Stability and Electronic Structure of GaN Nanotubes From Density-Functional Calculations. *Phys. Rev. B* **1999**, *60*, 7788.

(2) Rubio, A.; Corkill, J. L.; Cohen, M. L. Theory of Graphitic Boron Nitride Nanotubes. *Phys. Rev. B* **1994**, *49*, 5081.

(3) Pan, H.; Feng, Y. P.; Lin, J. Hydrogen Adsorption by Tungsten Carbide Nanotube. *Appl. Phys. Lett.* **2007**, *90*, 223104.

(4) Pan, H.; Feng, Y. P.; Lin, J. Ab Initio Study of Single-wall BC$_2$N Nanotubes. *Phys. Rev. B* **2006**, *74*, 045409.

(5) Menon, M.; Richter, E.; Mavrandonakis, A.; Froudakis, G.; Andriotis, A. N. Structure and Stability of SiC Nanotubes. *Phys. Rev. B* **2004**, *69*, 115322.

(6) Zhao, M.; Xia, Y.; Zhang, D.; Mei, L. Stability and Electronic Structure of AlN Nanotubes. *Phys. Rev. B* **2003**, *68*, 235415.

(7) Zhou, Z.; Zhao, J.; Chen, Z.; Gao, X.; Lu, J. P.; Schleyer, P. V. R.; Yang, C.-K. True Nanocable Assemblies with Insulating BN Nanotube Sheaths and Conducting Cu Nanowire Cores. *J. Phys. Chem. B* **2006**, *110*, 2529.

(8) Tang, C.; Bando, Y.; Huang, Y.; Yue, S.; Gu, C.; Xu, F.; Golberg, D. Fluorination and Electrical Conductivity of BN Nanotubes. *J. Am. Chem. Soc.* **2005**, *127*, 6552.

(9) Tang, C.; Bando, Y.; Ding, X.; Qi, S.; Golberg, D. Catalyzed Collapse and Enhanced Hydrogen Storage of BN Nanotubes. *J. Am. Chem. Soc.* **2002**, *124*, 14550.

(10) Kim, S. Y.; Park, J.; Choi, H. C.; Ahn, J. P.; Hou, J. Q.; Kang, H. S. X-ray Photoelectron Spectroscopy and First Principles Calculation of BCN Nanotubes. *J. Am. Chem. Soc.* **2007**, *129*, 1705.

(11) Suenaga, K.; Colliex, C.; Demoncy, N.; Loiseau, A.; Pascard, H.; Willaime, F. Synthesis of Nanoparticles and Nanotubes with Well-Separated Layers of Boron Nitride and Carbon. *Science* **1997**, *278*, 653.

(12) Wu, Q.; Hu, Z.; Wang, X.; Lu, Y.; Chen, X.; Xu, H.; Chen, Y. Synthesis and Characterization of Faceted Hexagonal Aluminum Nitride Nanotubes. *J. Am. Chem. Soc.* **2003**, *125*, 10176.

(13) Remskar, M. Inorganic Nanotubes. *Adv. Mater.* **2004**, *16*, 1497.

(14) Tenne, R.; Margulis, L.; Genut, M.; Hodes, G. Polyhedral and Cylindrical Structures of Tungsten Disulphide. *Nature* **1992**, *360*, 444.

(15) Feldman, Y.; Wasserman, E.; Srolovitz, D. J.; Tenne, R. High-Rate, Gas-Phase Growth of MoS_2 Nested Inorganic Fullerenes and Nanotubes. *Science* **1995**, *267*, 222.

(16) Hacohen, M. E.; Grunbaum, E.; Tenne, R.; Sloan, J.; Hutchison, J. L. Cage Structures and Nanotubes of $NiCl_2$. *Nature* **1998**, *395*, 336.

(17) Kaplan-Ashiri, I.; Cohen, S. R.; Gartsman, K.; Ivanovskaya, V.; Heine, T.; Seifert, G.; Wiesel, I.; Wagner, H. D.; Tenne, R. On the Mechanical Behavior of WS_2 Nanotubes Under Axial Tension and Compression. *PNAS* **2006**, *103*, 523.

(18) Rafailov, P. M.; Thomsen, C.; Gartsman, K.; Kaplan-Ashiri, I.; Tenne, R. Orientation Dependence of The Polarizability of An Individual WS_2 Nanotube by Resonant Raman Spectroscopy. *Phys. Rev. B* **2005**, *72*, 205436.

(19) Hamada, N.; Sawada, S.; Oshiyama, A. New One-dimensional Conductors: Graphitic Microtubules. *Phys. Rev. Lett.* **1992**, *68*, 1579.

(20) Pan, H.; Feng, Y. P.; Lin, J. First-principles Study of Optical Spectra of Single-wall BC_2N Nanotubes. *Phys. Rev. B* **2006**, *73*, 035420.

(21) Shubina, T. V.; Toropov, A. A.; Jmerik, V. N.; Tkachman, M. G.; Lebedev, A. V.; Ratnikov, V. V.; Sitnikova, A. A.; Vekshin, V. A.; Ivanov, S. V.; Kopev, P. S.; Bigenwald, P.; Bergman, J. P.; Holtz, P. O.; Monemar, B. Intrinsic Electric Fields in N-polarity GaN/$Al_xGa_{1-x}N$ Quantum Wells with Inversion Domains. *Phys. Rev. B* **2003**, *67*, 195310.

(22) Gudiksen, M. S.; Lauhon, L. J.; Wang, J.; Smith, D. C.; Lieber, C. M. Growth of Nanowire Superlattice Structures for Nanoscale Photonics and Electronics. *Nature* **2002**, *415*, 617.

(23) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.

(24) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of The Electron-gas Correlation Energy. *Phys. Rev. B* **1992**, *45*, 13244.

(25) Payne, M. C.; Teter, M. P.; Allan, D. C; Arias, T. A.; Joannopoulos, J. D. Iterative Minimization Techniques for ab initio Total-energy Calculations: Molecular Dynamics and Conjugate Gradients. *Rev. Modern Phys.* **1992**, *64*, 1045.

(26) Vanderbilt, D. Soft Self-consistent Pseudopotentials in A Generalized Eigenvalue Formalism. *Phys. Rev. B* **1990**, *41*, 7892.

(27) Monkhorst, H. J.; Pack, J. Special Points for Brillouin-zone Integrations. *Phys. Rev. B* **1976**, *13*, 5188.

(28) Teter, M. P.; Payne, M. C.; Allan, D. C. Solution of Schrödinger's Equation for Large Systems. *Phys. Rev. B* **1989**, *40*, 12255.

(29) Nepal, N.; Bedair, S. M.; El-Masry, N. A.; Lee, D. S.; Steckl, A. J.; Zavada, J. M. Correlation Between Compositional Fluctuation and Magnetic Properties of Tm-doped AlGaN Alloys. *Appl. Phys. Lett.* **2007**, *91*, 222503.

CT7003116

JCTC Journal of Chemical Theory and Computation

Density Functional and Basis Set Dependence of Hydrated Ln(III) Properties

Aurora E. Clark*

Department of Chemistry, Washington State University, P.O. Box 644630, Pullman, Washington 99164

Received November 20, 2007

Abstract: Benchmark studies of $Ln(H_2O)_{1,8-9}{}^{3+}$ (Ln = La, Lu) have been performed to assess the calculated properties obtained with local density approximation, generalized gradient approximation (GGA), meta-GGA, and hybrid functionals, when used with small- and large-core relativistic effective core potentials and their associated bases. Basis set dependence and the importance of specific functions to adequately describe the Ln atomic orbitals have been determined. The lanthanide contraction has been found to be an insufficient metric for characterizing the quality of a method/basis set combination due to cancellation of the errors. The electrostatic description obtained by natural population analysis has been examined, and an alternative partitioning of the valence space, which includes the 6s6p5d4f natural atomic orbitals, has been proposed.

Introduction

The aqueous chemistry of metal ions has an immense amount of literature available due to its fundamental importance in a variety of areas including biological systems, metal coordination, complexation behavior, and so forth. In the absence of metal-binding ligands, water molecules will coordinate to the metal to form the inner coordination shell. Within the alkali and alkaline earth series, the metal−water interaction is exclusively ionic in character.[1,2] In contrast, transition metal aquo complexes have coordination geometries that are strongly dependent upon the metal electronic state due to the participation of the d orbitals in the $M-OH_2$ bond.[3] Between these bonding extremes lie the lanthanide (Ln) and actinide (An) elements, whose f and d orbitals may participate in bonding with ligating water, but in a manner that is dependent upon the position of the element within the period and its oxidation state. In general, earlier An and Ln may have $M-OH_2$ interactions with some covalent character, while later in the series the interactions become more ionic.[4]

The solution chemistry of trivalent Ln has relevance to environmental remediation and the terrestrial migration of fission products at United States Department of Energy nuclear weapons facilities like the Hanford site. The potential closure of the nuclear fuel cycle currently proposes liquid/liquid extraction techniques for separating both 4f and 5f elements, mandating a fundamental understanding of the multiscale solution behavior of Ln and An. Using X-ray, extended X-ray absorption fine structure, and neutron diffraction methods, present experimental capabilities are able to determine bond lengths and angles that constitute molecular-scale solution geometry in a single phase from averaged ensembles, while long-range solvent organization is inferred from the bulk. Structural information within intermediate-length scales is experimentally formidable because it is buried in the bulk response of the condensed state. Owing to such experimental limitations, computational chemistry is a necessary complement to predict the molecular- and mesoscale solution behavior of f elements. Density functional theory (DFT) has been used for some time to calculate geometries and electronic structure of strongly bound Ln(III) complexes. Within that literature, it has been established that relativistic effective core potentials (RECPs) can capture the relativistic effects from all electron scalar-relativistic Douglas−Kroll−Hess calculations.[5] The relativistic effects upon the Ln−X bond lengths and the calculated Ln contraction have also been examined.[6–8] Both large- and small-core RECPs (and their associated basis sets) have been developed for Ln and are often used interchangeably, with little discussion of the influence of the f electrons upon electrostatic properties,

* Author e-mail: auclark@wsu.edu.

10.1021/ct700317p CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/05/2008

bonding, or the importance of basis sets upon calculated geometric and electronic structures. Comparisons of small- and large-core RECP[9,10] geometries do reveal longer Ln−X bond lengths from large-core calculations, presumably because of poorer treatment of the core-valence correlation relative to the small-core RECP. The extent to which the 4f electrons participate in bonding is a topic of current debate within the literature. While most studies indicate that the 4f orbitals/electrons do not participate in bonding, some systems (e.g., lanthanide trihalides) have shown pronounced 4f hybridization indicative of bonding interactions.[6]

In contrast to "strongly" bound Ln(III) complexes, the inner coordination sphere of aqueous Ln(III) is known to be dynamic, with significant exchange of the first- and second-shell H_2O being possible.[4] Lanthanides early in the series are predominantly nine-coordinate (nona-aqua) in solution, while late lanthanides have a propensity for being eight-coordinate (octa-aqua). Given the relative "weakness" of the $Ln-OH_2$ interaction, it is imperative that systematic benchmarks are performed to elucidate the most appropriate functionals, basis sets, and electronic structure analysis methods to be used. Indeed, such benchmarks may help elucidate patterns in geometric and electronic structure within the computational chemistry literature of strongly bound Ln systems. To this end, we have determined and analyzed the optimal geometric and electronic structures of $Ln(H_2O)_{1,8–9}{}^{3+}$ (Ln = La, Lu), using either local density approximation (LDA), generalized gradient approximation (GGA), meta-GGA, or hybrid density functionals in combination with small- and large-core RECPS and their associated contracted and uncontracted basis sets. A detailed analysis of the atomic orbital (AO) coefficients and atomic energies of Ln^{3+} cations has revealed specific functions necessary for describing the Ln AOs. Natural population analyses[11] with different partitioning of the core/valence/Rydberg natural atomic orbitals (NAOs) have also been studied, and a modified partitioning scheme has been proposed for calculating the Ln(III) charges. This series of molecules has been chosen due to their closed-shell electronic configurations, which allows for the basis set, density functionals, and RECPs to be examined in the absence of any errors associated with first-order spin–orbit coupling.

Computational Methods

The optimized structures of $Ln(H_2O)_{1,8–9}{}^{3+}$ (Ln = La, Lu) were obtained using local spin density approximation (LSDA; SVWN5 and the modern equivalent SPW92),[12–16] GGA (PBE, PW91, and B88P86),[17–22] hybrid (B3LYP and PBE0),[23–26] and meta-GGA functionals (TPSS).[27] These calculations were performed in NWChem[28] and Gaussian03.[29] The former employed a self-consistent field (SCF) energy convergence criterion of 10^{-6}, an integral internal screening threshold of 10^{-16}, a numerical integration grid of 10^{-8}, and a tolerance in Schwarz screening for the Coulomb integrals of 10^{-12}. Gaussian03 calculations used the Ultrafine integration grid (99 590 points), and SCF convergence was set to "verytight" (10^{-6}). "Verytight" optimization convergence was not used in all cases due to computational expense. In test cases, differences in energy between the default and "verytight" optimization criteria were less than 1.6 millihartrees. All geometries were confirmed to be local minima, with no imaginary vibrations unless otherwise noted. Both small-core Stuttgart−Dresden (SD) RECP[30,31] (which includes the $n = 4–6$ shells in the valence space) and large-core SD RECP[32] (which includes the $n = 5$ and 6 shells in the valence space) were examined with their associated generally contracted, segmented contracted, and uncontracted basis sets. The oxygen and hydrogen atoms were treated with an aug-cc-PVDZ basis set.[33] Natural population[11] and Mulliken population[34] analyses were performed at the optimized geometries. All calculations were performed on the massively parallel Linux cluster in the Molecular Science Computing Facility in the William R. Wiley Environmental Molecular Sciences Laboratory at the Pacific Northwest National Laboratory, or at the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science user facility at Lawrence Berkely National Laboratory.

Results and Discussion

It is well-known that the lanthanide series exhibits a pronounced decrease in the ionic radii with increasing atomic number.[35] Filling the 4f orbitals improves shielding of the nuclear charge and is most pronounced early in the series. Subsequently, the decrease in ionic radii is larger at the beginning of the series than at the end. This trend may be monitored by examining the monotonic decrease of Ln−X bond lengths (X = Lewis-base donor), wherein a quadratic dependence across the series has been observed. This behavior is observed in X-ray structures of isostructural Ln complexes[36,37] (Ln = La–Lu), in addition to more limited sets of solid-state materials[38,39] and coordination compounds.[40,41] Quadrelli suggested that individual classes of bond lengths can be fit by a second-order polynomial.[36] More recent studies by Seitz et al.[37] have indicated that the ligand field responds to a change in the average metal ion size to distribute the metal–ligand bond-length changes; however, taking the average bond length does show the anticipated contraction. There, it was pointed out that the quadratic dependence of the lanthanide contraction can be derived from the model proposed by Slater[42] and later modified by others.[43,44] This model utilizes empirical rules for the shielding of the nuclear charge Z from electrons in a particular orbital by inner electron shells, expressed by a screening constant s.

The lanthanide contraction has been calculated by Pyykko[45] and others through the comparison of the difference of the ionic radii of lanthanum and lutetium, as measured by bond length Ln−X for isostructural species:

$$\Delta_{Ln} = r_e(LaX) - r_e(LuX)$$

Experimentally and theoretically, the Ln contraction has been found to be dependent upon the coordination number, the charge of the ions, and, some have suggested, bond type. Generally, large contractions are observed for soft bonds and small contractions are observed for stiffer ones.[46] Importantly, the calculated Δ_{Ln} is often used as a metric for assessing the reliability and quality of an *ab initio* calculation.

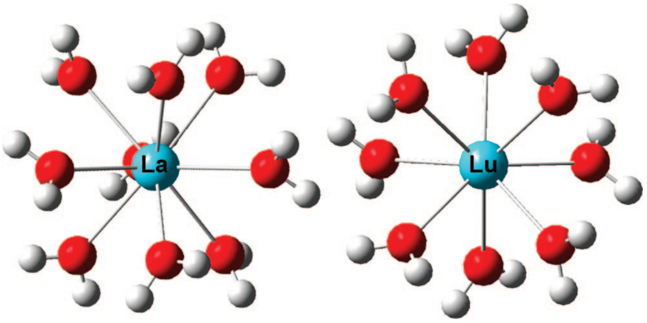


Figure 1. DFT-optimized tricapped trigonal bipyramidal $La(H_2O)_9{}^{3+}$ and square antiprismatic structure of $Lu(H_2O)_8{}^{3+}$. Specific geometric parameters are presented in Tables 1 and 2.

This is problematic, as computationally important errors within calculations on La and Lu may cancel when determining the contraction. Here, we use a variety of metrics to assess the quality of particular method/basis set combinations, illustrating the potential pitfalls of relying on computed lanthanide contraction values.

Geometric Dependence Upon RECP and Basis. *Small-Core RECP and Segmented Contracted Basis.* Two small-core RECPs (28 e^- in the core) and basis sets are available for La^0 and Lu^0 from the Stuttgart group. The first basis set to be considered is the (14s13p10d8f6g)/[10s8p5d4f3g], which is based upon a segmented contraction scheme.[31] Here, the most diffuse s function, which has an exponent of 0.02 in La^0 and 0.03 in Lu^0, has been removed to yield a [9s8p5d4f3g] contracted Gaussian basis, so as to prevent linear dependence errors. The performance of the (13s13p10d8f6g)/[9s8p5d4f3g] basis was first examined by B3LYP optimization of $La(H_2O)_9{}^{3+}$ and $Lu(H_2O)_8{}^{3+}$ in the gas phase and subsequent comparison with available experimental data. As seen in Figure 1, the nona-aqua La(III) adopts a tricapped trigonal bipyramidal structure, while the octa-aqua Lu(III) adopts a square antiprismatic arrangement of the ligating waters. X-ray structures in dilute $LaCl_3$ solutions observe an average nona-aqua $La-OH_2$ bond distance of 2.580 Å, while dilute $LuCl_3$ solutions have average octa-aqua $Lu-OH_2$ distances of 2.338 Å.[47,48] This compares well (within 0.03 Å) with the average calculated B3LYP values of $r_{La-OH_2} = 2.618$ Å and $r_{Lu-OH_2} = 2.369$ Å. IR data are also available for dilute solutions of $[Ln(H_2O)_{9-8}{}^{3+}]$ $(C_2H_5SO_4)_3$ wherein the $Ln-OH_2$ stretch for $La(H_2O)_9{}^{3+}$ occurs at 316 cm^{-1} and, for $Lu(H_2O)_8{}^{3+}$, at 342 cm^{-1}.[49] B3LYP calculates these stretches to be at 279 and 329 cm^{-1} (unscaled), respectively (Table 1). These low vibrational frequencies are consistent with the somewhat too long $Ln-OH_2$ bond lengths.

To examine the lanthanide contraction, Δ_{Ln}, the $Ln-OH_2$ bond lengths for isostructures $La(H_2O)_8{}^{3+}$ and $Lu(H_2O)_8{}^{3+}$ were compared (Figure 1 and Table 1). The calculated lanthanide contraction for these hydrates is 0.215 Å, in excellent agreement (±0.005 Å) with four-component relativistic Hartree−Fock (HF; DHF) and MP2 (RMP2) calculations of $Ln(H_2O)^{3+}$.[50] Given previous observations regarding the sensitivity of the lanthanide contraction to the coordination number and type of ligands, we sought to directly compare the DHF and RMP2 values with B3LYP by calculating the optimized geometries of $La(H_2O)^{3+}$ and $Lu(H_2O)^{3+}$. Here, the hybrid functional predicts $r_{La-OH_2} = 2.289$ Å and $r_{Lu-OH_2} = 2.094$ Å. This compares to reported DHF values of 2.34 and 2.13 Å, respectively, and RMP2 values of 2.29 and 2.08 Å, respectively. Thus, there is clearly excellent agreement between B3LYP (using a RECP basis) and RMP2 for individual bond lengths, and all three methods yield lanthanide contraction values that deviate by only 0.01 Å from each other.

The number of valence functions within the contracted segmented set is quite large, particularly when compared to the small-core contracted basis sets for An that are routinely used (12s11p10d8f)/[8s7p5d4f] and truncated to [6s6p5d3f].[51,52] Consequently, the importance of the most diffuse functions to the total electronic energy, geometry, IR frequencies, and lanthanide contractions of isostructural $La(H_2O)_8{}^{3+}$ and $Lu(H_2O)_8{}^{3+}$ was examined. The most diffuse functions from each shell were first systematically removed and the single-point energies calculated at the optimized [9s8p5d4f3g] geometries. The energy differences between the original and truncated bases are indicative of the contribution of the omitted basis to the total electronic energy. The geometry was then reoptimized using the truncated basis to determine the importance of the omitted functions to structural parameters and the frequency of the IR-active totally symmetric $Ln-OH_2$ stretch, $\nu_{Ln-H_2}O$ (Table 1).

The first truncated basis omits the three g functions to yield the [9s8p5d4f] set. The energetic consequence of the g functions is minor, as the total electronic energy from the single-point calculation is increased by only 1.6 and 2.4 millihartrees for octa-aqua La(III) and Lu(III), respectively. The optimization of $Ln(H_2O)_8{}^{3+}$ with the [9s8p5d4f] basis yields nearly identical geometries and vibrational frequencies to those obtained with [9s8p5d4f3g]. Omission of the most diffuse f function, which has an exponent of 0.1973 for La^0, increases the single-point energy of $La(H_2O)_8{}^{3+}$ by 3.4 millihartrees, and subsequent geometry optimization implies minimal impact on the structure and IR spectra. Yet, removal of the most diffuse f function on Lu^{3+} (exponent of 0.4244) increases the single-point electronic energy of $Lu(H_2O)_8{}^{3+}$ by 108.7 millihartrees in the [9s8p5d3f] basis. Subsequently, geometry optimization decreases the $Lu-OH_2$ bond length by 0.05 Å and shifts the frequency of the symmetric stretch, ν_{Lu-OH_2}, by 30 cm^{-1} relative to that calculated in the larger [9s8p5d4f] basis. Truncating the most diffuse p function for La^0 and Lu^0 (with a 0.08 exponent), to yield the [9s7p5d43f] basis, alters the single-point electronic energies by less than 1.6 millihartrees and has negligible structural consequences. However, removing the second most diffuse p functions with exponents of 0.2292 and 0.2858 on La^0 and Lu^0, respectively, increases the energy by 63.4 and 38.9 millihartrees. The net affect after geometry optimization in the [9s6p5d3f] basis is a contraction of r_{La-OH_2} by 0.06 Å and a 50 cm^{-1} shift in ν_{La-OH_2} in $La(H_2O)_8{}^{3+}$, and a 0.03 Å bond-length shortening in $Lu(H_2O)_8{}^{3+}$ with a 30 cm^{-1} shift in ν_{Lu-OH_2}. Turning to the s functions, we observe that truncating the most diffuse s, with exponents of 0.0467 and 0.079 on La^0 and Lu^0, has little affect upon the energy (<1.5 millihartree) or geometry; however, omission of the functions with La^0 and Lu^0

Table 1. C_1 B3LYP Structural Parameters and the Symmetric Ln−OH$_2$ Vibrational Stretching Frequency (cm^{-1}), $\nu_{\mathrm{Ln-OH_2}}$, Obtained for La(H$_2$O)$_{8-9}{}^{3+}$ and Lu(H$_2$O)$_8{}^{3+}$ as a Function of General (ano) and Segmented (seg) Contracted Metal Basis Sets Using Both Small-Core (ECP28MWB) and Large-Core (ECP47MWB for La^{3+}, ECP60MWB for Lu^{3+}) RECPs[a]

	~symm	tol	$<r_{\mathrm{M-O}}>$	$\theta_{\mathrm{O1-M-O2}}$	$\nu_{\mathrm{Ln-OH_2}}$[b,c]	E	$<\Delta\mathrm{Ln}>$
			$La(H_2O)_9{}^{3+}$				
ECP28MWB							
[9s8p5d4f3g] seg	C_3	0.01	2.618	69.7	279	−1122.987628	
[9s8p5d3f] seg	C_3	0.01	2.620	69.8	280	−1122.981900	
			$La(H_2O)_8{}^{3+}$				
ECP28MWB							
[6s6p5d4f3g] ano	S_8, C_4, C_2	0.010	2.580	78.0	293	−1046.505984	0.212
[9s8p5d4f3g] seg	S_8, C_4, C_2	0.010	2.580	78.0	293	−1046.507198	0.215
[9s8p5d4f] seg	S_8, C_4, C_2	0.010	2.582	78.0	293	−1046.504786	0.213
[9s8p5d3f] seg	S_8, C_4, C_2	0.010	2.581	78.0	294	−1046.501456	0.254
[9s7p5d3f] seg	S_8, C_4, C_2	0.010	2.581	78.0	294	−1046.501092	0.255
[9s6p5d3f] seg	C_4, C_2	0.001	2.525	76.6	342	−1046.441367	0.223
[8s6p5d3f] seg	C_4, C_2	0.001	2.524	76.6	342	−1046.440913	0.224
[7s6p5d3f] seg	C_4	0.001	2.506	74.9	358	−1046.417898	0.240
[6s6p5d3f] seg	S_8, C_4, C_2	0.010	2.332	69.4	565	−1046.055109	0.275
ECP47MWB							
[5s4p3d]	S_8, C_4, C_2	0.010	2.603	78.0	288	−642.433080	0.213
			$Lu(H_2O)_8{}^{3+}$				
ECP28MWB							
[6s6p5d4f3g] ano	S_8, C_4, C_2	0.100	2.369	77.9	328	−1847.144368	
[9s8p5d4f3g] seg	S_8, C_4, C_2	0.100	2.365	77.8	329	−1847.129760	
[9s8p5d4f] seg	C_2	0.010	2.369	77.9	329	−1847.128037	
[9s8p5d3f] seg	S_8, C_4, C_2	0.010	2.327	78.0	353	−1847.021075	
[9s7p5d3f] seg	S_8, C_4, C_2	0.010	2.326	78.0	354	−1847.020175	
[9s6p5d3f] seg	S_8, C_4, C_2	0.010	2.301	77.6	384	−1846.985048	
[8s6p5d3f] seg	S_8, C_4, C_2	0.010	2.300	77.6	385	−1846.984341	
[7s6p5d3f] seg	C_2	0.001	2.266	75.1	436	−1846.926420	
[6s6p5d3f] seg	C_1		2.057	73.0	629	−846.201021	
ECP60MWB							
[5s4p3d]	S_8, C_4, C_2	0.010	2.390	77.9	322	−650.797635	

[a] Nearest symmetry group (symm), tolerance to reach higher symmetry (tol), average bond lengths ($<r_{\mathrm{M-O}}>$ in Å), maximum deviation from the average bond lengths ($\Delta r_{\mathrm{M-O}}{}^{\mathrm{max}}$ in Å), bond angles ($\theta_{\mathrm{O1-M-O2}}$ in deg), total electronic energies (E in hartrees), and average Ln contraction ($<\Delta\mathrm{Ln}>$ in Å) are presented. [b] Experimental value = 316 cm^{-1}, ref 49. [c] Experimental value = 342 cm^{-1}, ref 49.

exponents of 0.2539 and 0.4408 increases the energy by up to 82.9 millihartree. Similar to the above observations, removal of the second most diffuse s function shortens the metal–ligand bond lengths by ~0.04 Å and increases the energy of the symmetric metal−water vibrational stretch by ~50 cm^{-1}. Completely unrealistic energies, geometries, and frequencies are obtained with the [6s6p5d3f] basis.

Despite the dramatic changes that occur in geometry when using these truncated basis sets, it is quite interesting to note that the value of lanthanide contraction is relatively constant. Indeed, Δ_{Ln} varies by only 0.025 Å between the [9s8p5d4f3g] and [7s6p5d3f] bases, clearly indicating that the lanthanide contraction is a poor metric for assessing the quality of geometries of lanthanide complexes. When these structural and vibrational data are compared as a function of the basis with available experimental data, the closest agreement for Lu(H$_2$O)$_8{}^{3+}$ occurs with the [9s8p5d3f] and [9s7p5d3f] basis sets. Since the largest basis set does not yield the closest geometric parameters relative to experimental results, there are obvious methodological errors associated with using the B3LYP functional (*vide infra*). However, a second solvation shell may also improve the agreement with experimental structures and frequencies, creating more consistency in the accuracy of the various basis sets. While it might be tempting to assume that the same truncated basis sets that perform well for octa-aqua Lu(III) would yield similar quality results for nona-aqua La(III), geometry optimization of La(H$_2$O)$_9{}^{3+}$ with either the [9s8p5d3f] or [9s7p5d3f] basis sets yields structures that deviate farther from experiment than that obtained with the [9s8p5d4f] basis (Table 1). As a consequence of the energetic and geometric observations as a function of the basis, those bases smaller than [9s8p5d4f] are not recommended for use when studying the solvation properties of Ln(III).

This brings up the question, however, of why the truncated basis sets perform so badly. Contraction errors and inflexibility of the basis may be responsible, or perhaps the truncated functions are genuinely necessary for an accurate description of the Ln atomic orbitals. Intertwined is the fact that DFT functionals incorporate some amount of correlation energy, which may itself lead to deviations in the treatment of contracted and uncontracted bases, and it is further possible that the DFT density may be significantly different from the Hartree−Fock (HF) density for which the basis set was developed. To explore these issues, the total electronic energies and AO coefficients of La^{3+} and Lu^{3+} cations were systematically examined using HF, a single SCF cycle of DFT fixed at the HF density, and SCF-optimized DFT. Both the segmented contracted and uncontracted bases were investigated, and we define the difference in energy between the two bases as the contraction error:

$$\Delta E_{\mathrm{cont}} = E_{\mathrm{contracted}} - E_{\mathrm{uncontracted}}$$

Table 2. Structural Parameters and the Symmetric Ln−OH$_2$ Vibrational Stretching Frequency (cm^{-1}), ν_{Ln-OH_2}, of Ln(H$_2$O)$_{8-9}$$^{3+}$ (Ln = La, Lu) Obtained with LDA, GGA, and meta-GGA Functionals Using the Small-Core RECP and the [9s8p5d4f3g] Segmented Basis[a]

method	∼symm	tol	$<r_{M-O}>$	$\theta_{O1-M-O2}$	ν_{Ln-OH_2}	$<\Delta Ln>$
			La(H$_2$O)$_9$$^{3+}$			
TPSS	C_3	0.01	2.600	69.8	280	
B3LYP	C_3	0.01	2.618	69.7	279	
			La(H$_2$O)$_8$$^{3+}$			
SVWN5	S_8, C_4, C_2	0.010	2.496	78.0	326	0.213
SPW92[b]	C_2	0.001	2.496	78.0	321	0.213
PW91	S_8, C_4, C_2	0.010	2.564	77.9	294	0.208
B88P86	S_8, C_4, C_2	0.010	2.569	78.0	291	0.209
PBE	S_8, C_4, C_2	0.010	2.569	78.0	292	0.207
TPSS	S_8, C_4, C_2	0.010	2.565	78.0	294	0.214
B3LYP	S_8, C_4, C_2	0.010	2.580	78.0	293	0.215
PBE0	S_8, C_4, C_2	0.010	2.556	77.8	300	0.191
			Lu(H$_2$O)$_8$$^{3+}$			
SVWN5	C_2	0.010	2.283	77.5	368	
SPW92	C_4, C_2	0.010	2.283	77.7	351	
PW91	S_8, C_4, C_2	0.010	2.356	77.8	328	
B88P86	S_8, C_4, C_2	0.010	2.360	77.9	325	
PBE	S_8, C_4, C_2	0.010	2.361	77.8	325	
TPSS	S_8, C_4, C_2	0.010	2.351	77.6	327	
B3LYP	S_8, C_4, C_2	0.100	2.365	77.8	329	
PBE0	S_8, C_4, C_2	0.100	2.365	77.8	322	

[a] Nearest symmetry group (symm), tolerance to reach higher symmetry (tol), average bond lengths ($<r_{M-O}>$ in Å), maximum deviation from the average bond lengths (Δr_{M-O}^{max} in Å), bond angles ($\theta_{O1-M-O2}$ in deg), total electronic energies (Hartree), and average Ln contraction ($<\Delta Ln>$ in Å) are presented. [b] Optimized structure had a single imaginary vibration at −33 cm^{-1}.

Using HF, ΔE_{cont} values for the [9s8p5d4f3g] segmented basis of La^{3+} and Lu^{3+} are 0.34 and 4.51 millihartrees, respectively. B3LYP performed at the HF density with a single SCF cycle yields contraction errors for La^{3+} and Lu^{3+} of 0.24 and 1.16 millihartrees, indicating that the intrinsic correlation obtained by B3LYP should not affect basis set performance relative to HF. In contrast, the correlation error obtained from the SCF-optimized B3LYP energies of La^{3+} and Lu^{3+} are 0.66 and 24.0 millihartrees, respectively. This indicates that, while the B3LYP density for La^{3+} is quite similar to that of HF, it is significantly different for Lu^{3+}. These numerical results are shown pictorially in Figure 2, along with plots of the uncontracted HF density, ρ_{HF}, and the B3LYP difference densities ($\rho_{HF} - \rho_{B3LYP}$) for La^{3+} and Lu^{3+}. In the case of La^{3+}, B3LYP and HF both calculate the same density, and thus ($\rho_{HF} - \rho_{B3LYP}$) is negligible everywhere (Figure 2a). However, for Lu^{3+}, it is clear that the B3LYP density deviates significantly relative to that obtained by HF (Figure 2b). Thus, the large contraction error obtained from the B3LYP calculation on Lu^{3+} (Figure 2, right-hand panel) arises from the fact that the DFT density using the contracted basis is very similar to that of HF, while uncontracting the basis leads to a different density and energy (Figure 2b).

Comparison of the HF and B3LYP SCF AO coefficients for Lu^{3+} indicates that, in general, the more diffuse basis functions have larger contributions to the occupied AOs in B3LYP than in HF, which leads to the density difference shown in Figure 2b. Monitoring the HF and B3LYP energies for Lu^{3+} with the uncontracted segmented basis as a function of truncation level reveals significant changes when specific functions are removed from the basis, indicating that HF calculations will experience the same basis set dependence as B3LYP. The energetically crucial functions are the 10th to 12th s, having exponents of 0.4408, 1.0287, and 2.6778, respectively, and the 12th p in the (13s13p10d8f6g), which has an exponent of 0.2858. This is the same behavior observed in Lu(H$_2$O)$_8$$^{3+}$, illustrating that these functions are crucial for the correct description of the atomic orbitals of Lu(III). The HF and B3LYP SCF AO coefficients show that these functions help describe the 4s, 5s, and 5p AOs within the [Ar]4s^{2}3d^{10}4p^{6}5s^{2}4d^{10}5p^{6}4f^{14} electronic configuration of Lu^{3+}, where the [Ar]3d^{10} electrons are in-core. Interestingly, the most diffuse uncontracted f function (with an exponent of 0.4244) has a significant energetic consequence, yet it participates only in the unoccupied f atomic orbitals. The importance of diffuse functions is not necessarily surprising, as such functions are likely needed to describe its filled 4f shell. Monitoring the total HF and B3LYP electronic energies for La^{3+} ([Ar]4s^{2}3d^{10}4p^{6}5s^{2}4d^{10}5p^6 electronic configuration) using the uncontracted basis as a function of truncation reveals significant changes to both the HF and B3LYP energies when the p function with an exponent of 0.2292 is removed (the ninth p in the (13s13p10d8f6g) set), and similarly when the s functions with 0.5672 and 0.2539 exponents are removed (the 11th and 12th s functions in the (13s13p10d8f6g) set, respectively). The SCF basis function coefficients indicate that the 11th s function contributes to the 4s and 5s AOs, while the 12th s function has large coefficients for the 5s AO, and the ninth p function contributes to the 4p and 5p AOs. On the basis of these results, it is clear that truncating the Stuttgart small-core basis set amounts to removing key functions necessary to describe the Ln atomic orbitals.

Small-Core RECP and General Contracted Basis. The second basis set to be considered is the (14s13p10d8f6g)/[6s6p5d4f3g] atomic natural orbital basis set, which is based upon a generalized contraction scheme.[29] These geometries (Table 1) are found to be nearly identical to those obtained by the segmented [9s8p5d4f3g]. The agreement between the two bases is slightly better in La(H$_2$O)$_8$$^{3+}$ (r_{La-OH_2} deviations of 0.0001 Å) than in Lu(H$_2$O)$_8$$^{3+}$ (r_{Lu-OH_2} deviations of 0.003 Å). This result is not entirely unanticipated, as the B3LYP La^{3+} and Lu^{3+} cation densities are nearly identical using either the uncontracted segmented or generalized contracted natural orbital basis.

Large-Core RECP and Basis. Previous calculations[9] utilizing large-core RECPs for Ln have noted that placement of the f electrons and orbitals in the core leads to qualitatively similar structural parameters, but with increased bond lengths. This is presumably due to poorer treatment of core−valence correlation relative to the small-core RECPs. The large-core Stuttgart RECP for La^{3+} places the [Kr]4d^{10} electrons in the core, leaving 5s^{2}5p^6 in the valence space. Geometry optimization of La(H$_2$O)$_8$$^{3+}$ yields a square antiprismatic structure near S_8 symmetry, similar to those found using the two small-core RECPs. The average La−OH$_2$ bond length is 2.603 Å, which is 0.023 Å longer than the small-core bond lengths using the (13s13p10d8f6g)/[9s8p5d4f3g] segmented

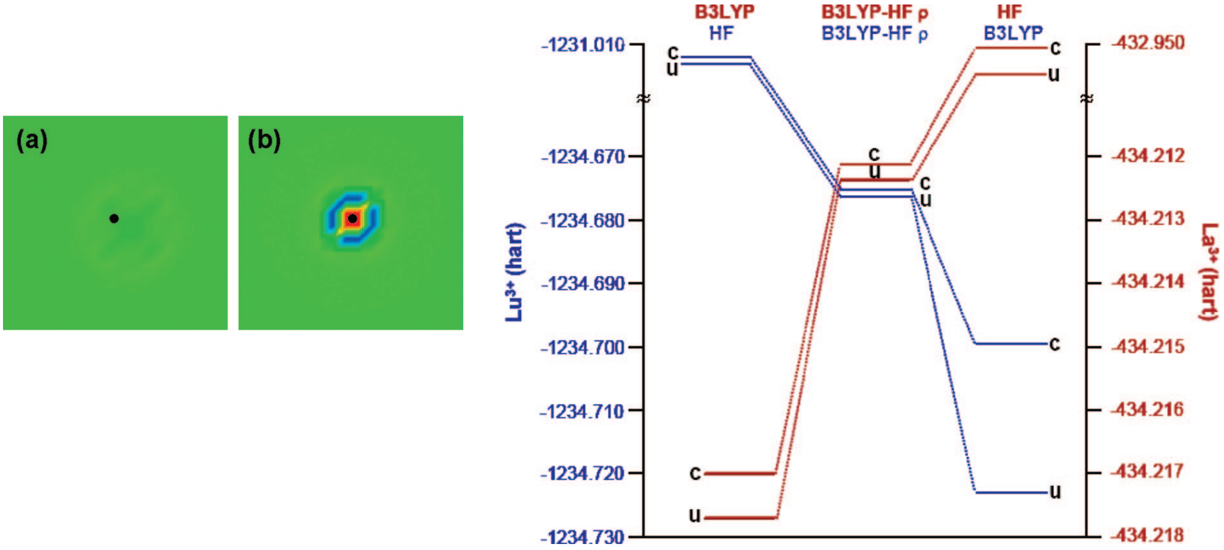


Figure 2. Left-hand panel: the difference density between HF and B3LYP ($\rho_{HF} - \rho_{B3LYP}$) for (a) La^{3+} and (b) Lu^{3+} using the uncontracted small-core [9s8p5d4f3g] basis [scale of the density difference: –6.79 × 10^{-2} (blue), 7.23 × 10^{-2} (red)]. Right-hand panel: the total electronic energies of La^{3+} (red) and Lu^{3+} (blue) with HF, B3LYP fixed at the HF density, and fully SCF-optimized B3LYP using the small-core [9s8p5d4f3g] contracted (c) and uncontracted (u) basis sets.

and (14s13p10d8f6g)/[6s6p5d4f3g] general contracted basis sets, respectively. The total number of H bonds present in $La(H_2O)_8{}^{3+}$ is the same in all three cases; however, the large-core basis also increases the average H-bond length from ~2.958 Å in the small-core calculations to 2.986 Å. Similar results are obtained for $Lu(H_2O)_8{}^{3+}$, wherein the $Lu-OH_2$ bond length increases by 0.025 Å and the average H-bond length for the 16 hydrogen bonds increases by 0.032 Å relative to the small-core (13s13p10d8f6g)/[9s8p5d4f3g] segmented calculation. This is important, as it indicates that the large-core RECP affects not only the immediate Ln−X bond length but also structural features that extend beyond the metal's nearest bonding interactions. Interestingly, the perturbations in hydrogen bonding as a function of RECP are not attributable to differences in the electrostatic interactions, as both Mulliken[34] and NPA[11] charges are virtually unchanged (*vide infra*).

Geometric Dependence Upon Density Functionals. The Jacob's ladder of density functionals[53] describes the relative increase in the ability of approximate exchange and correlation functionals, E_{XC}'s, to approach the density given by the exact many-body wave function. Depending on the level of covalency within the metal–ligand bond, LDA, GGA, meta-GGA, and hybrid functionals can give quite different electronic and geometric descriptions for f-element complexes.[54] Using LDA with spin–orbit corrections on lanthanide atoms has been shown to yield ionization potentials close to experimental values,[55] yet the electronic description becomes less accurate in molecular calculations, and overbinding is often predicted.[56,57]A variety of GGA functionals (BLYP, BP, BPW, and PWPW) have been successful in describing LnX_3 compounds (X = F, Cl, Br, I; Ln = La, Gd, Lu),[55] utilizing the RECPs and basis sets of Cundari and Stevens.[58,59] Some studies have shown that hybrid functionals predict significantly longer bond lengths than GGA functionals,[54] yet B3LYP is often the functional of choice within computational lanthanide chemistry, and very successful studies have been performed.[9] In order to assess the applicability of different approximate exchange correlation functionals to describe hydrated Ln(III), we have examined the geometries and lanthanide contraction values of isostructural $La(H_2O)_8{}^{3+}$ and $Lu(H_2O)_8{}^{3+}$ using LSDA, GGA, meta-GGA, and hybrid functionals with the small-core RECP and the [9s8p5d4f3g] segmented basis (Table 2).

In general, each density functional predicts similar square antiprismatic geometries for $Ln(H_2O)_8{}^{3+}$. Most have a high degree of symmetry and are near S_8, though $Lu(H_2O)_8{}^{3+}$ is typically less symmetric than $La(H_2O)_8{}^{3+}$. For the sake of comparison to experimental data, the functional performance for $Lu(H_2O)_8{}^{3+}$ will first be discussed. Recall that, experimentally, $r_{Lu-OH_2} = 2.338$ Å and $\nu_{Lu-OH_2} = 342$ cm^{-1}. Given prior observations of overbinding in LDA,[56,57] it is not surprising that LSDA predicts the shortest metal–ligand bond lengths, $r_{Lu-OH_2} = 2.283$ Å. Interestingly, its calculated frequency for the symmetric metal–ligand stretch shifts by 17 cm^{-1}, depending upon the functional, with SPW92 predicting the closest value, $\nu_{Ln-OH_2} = 351$ cm^{-1}, relative to experimental results. Within the GGA functionals, PW91 yields metal–ligand bond lengths that are ~0.04 Å shorter than those of B88P86 and PBE; yet, each GGA predicts very close ν_{Ln-OH_2} values ($\nu_{Ln-OH_2} = 325-328$ cm^{-1}). The meta-GGA TPSS functional agrees most closely with the experimental average $Lu-OH_2$ distance (within 0.013 Å); however, its calculated frequency for the symmetric metal–ligand stretch is comparable to the GGA and hybrid functionals, all of which predict too low of a value by ~20 cm^{-1}. The hybrid B3LYP and PBE0 functionals predict the longest r_{Lu-OH_2} values, overestimating the distance by 0.03 Å.

Additionally, some of the trends in structure for $Lu(H_2O)_8{}^{3+}$ do not hold for the octa-aqua La(III) complex. For example, there is a 0.02 Å deviation in the $La-OH_2$ bond length between the B3LYP and PBE0 hybrid functionals in $La(H_2O)_8{}^{3+}$, while the calculated $Lu-OH_2$ bond lengths for the same functionals are essentially identical. Indeed, the hybrid PBE0 functional yields a metal–ligand bond length for $La(H_2O)_8^{3+}$ that is much closer to that predicted by TPSS than that by B3LYP. Examining the isostructural octa-aqua species allows for investigation of the functional dependence of the calculated lanthanide contraction values. LSDA, TPSS, and B3LYP predict essentially the same Δ_{Ln} of 0.21 Å, while the GGA and PBE0 functionals predict somewhat shorter values. This again highlights the misleading nature of the lanthanide contraction as a metric for assessing the appropriateness of a given method, as techniques with nearly identical Δ_{Ln} values may have quite different individual metal–ligand bond lengths that may or may not agree well with experimental values. Given the good performance of the TPSS functional for $Lu(H_2O)_8^{3+}$, the structure of nona-aqua La(III) was optimized and those results compared to experimental ones (Table 2). Indeed, TPSS does yield slightly better agreement with the X-ray average r_{La-OH_2} than B3LYP (by 0.01 Å), yet it has essentially the same predicted frequency for the totally symmetric La−water stretch, which is nearly 40 cm^{-1} below the reported value.[49] It thus appears that no functional performs equally well for calculated bond lengths as for IR frequencies in the hydrated species with a single solvent shell. Some improvement may be expected upon increasing the basis set on the water ligands, or by adding a second solvation shell. Another experimental observable of interest to be compared is the free energy of hydration for Ln(III). That topic is beyond the scope of this paper, as it is also highly dependent upon the solvation models used; however, it is the topic of another manuscript.[60]

Population Analyses. Population analyses are a common way to characterize the electronic structure of metal atoms. Often, the large basis sets needed to describe d- and f-block metals makes standard Mulliken[34] and Löwdin[61] methods nonoptimal. Natural population analysis[11] (NPA) has emerged as one of the methods of choice for such systems because its partitioning of the orbital space decreases basis set dependence. NPA divides the molecular charge into atomic components in analogy to Löwdin's method,[61] where an occupancy-weighted symmetric (Löwdin) transformation is used to select an AO and partition it to a set of orbitals labeled "core," "valence", and "Rydberg," each of which contributes differently to the density. Core orbitals contribute exactly 2*e* to the atomic population, while the valence orbitals vary in their contributions and the Rydberg set participates minimally in the charges. In previous studies, our and other groups have highlighted the sensitivity of NPA charges to the initial partitioning of the NAO basis into valence and Rydberg sets. This is particularly true with regard to metals, where in the d block the default NPA partitioning of the NAOs excludes the formally empty set of p orbitals from the valence space of the metal atom. Including the 4p AO in the Rydberg set can lead to a larger positive charge on the metal than if it is considered part of the valence space, owing to the interaction between the empty p and the ligand orbitals.[62] The correct partitioning of the NAOs also influences predicted trends in atomic charges within a series. In our studies of 5f actinide complexes, we noted that the empty set of 6d orbitals (which may contribute to the bonding of the actinyls) is placed in the Rydberg basis and not in the valence. Altering the default partitioning scheme to have the 7s5f6d in the valence successfully reproduced trends in electron-donating capability within the equatorial ligands bound to UO_2^{2+}.[63]

Table 3. NPA Charges as a Function of Valence/Rydberg Partition of the NAO Basis[a] for $La(H_2O)_8^{3+}$ and $Lu(H_2O)_8^{3+}$ Using the Small-Core RECP and B3LYP/[9s8p5d4f3g]/aug-cc-pvdz

Valence/Rydberg partition	q_{Ln}
$La(H_2O)_8^{3+}$	
0/6–10s; 6–10p; 5–8d; 4–6f	2.846
6s/7–10s; 6–10p;5–8d; 4–6f	2.768
6s; 4f/7–10s; 6–10p; 5–8d; 5–6f	2.695
6s; 6p/7–10s; 7–10p; 5–8d; 4–f	2.542
6s; 6p; 5d/7–10s; 7–10p; 6–8d; 4–6f	2.032
6s; 6p; 5d; 4f/7–10s; 7–10p; 6–8d; 5–6f	1.948
$Lu(H_2O)_8^{3+}$	
0/6–11s; 6–10p; 5–8d; 5–7f	2.848
6s/7–11s; 6–10p; 5–8d; 5–7f	2.681
6s; 6p/7–11s; 7–10p; 5–8d; 5–7f	2.350
6s;5d/7–11s;7–10p;6–8d;5−7f	2.191
6s; 6p; 5d/7–11s; 7–10p; 6–8d; 5–7f	1.857

[a] The core NAOs of La(III) are the 4–5s; 4–5p; 4d and for Lu(III) are 4–5s; 4–5p; 4d; 4f. Default partitioning schemes are in bold.

In a similar vein, we have examined the appropriate NAOs to be placed in the valence space of trivalent La and Lu as well as the dependence of NPA charges upon the functional and basis set. This is particularly important as NPA has been extensively used to understand a variety of electronic effects including the influence of higher coordination numbers upon energetics of Ln(III) reactions and electronic structure.[9,64,65] The default partitioning for La(III) in $La(H_2O)_8^{3+}$ places the 4s and 5s, 4p and 5p, and 4d NAOs in the core; the 6s and the 4f in the valence; and the 7–10s, 6–10p, and 5f and 6f in the Rydberg space. To test the calculated metal charge as a function of the valence/Rydberg NAO partition, all valence NAOs were first placed in the Rydberg set and then systematically brought into the valence space (Table 3). In the case of zero NAOs participating in the valence set, the atomic charge on La(III) is quite near the formal charge of 3+ ($q_{La} = 2.846$). Inclusion of the 6s NAO in the valence space decreases the charge by 0.08*e*, and bringing in the 4f NAO adds another 0.07*e* to La(III). Thus, the charge on La(III) using the default NPA partitioning is 2.695. Similar contributions to the atomic charge are found with the 6p orbital in the valence; however, the largest effect is by far observed when the 5d NAO is allowed to participate in the valence. The 5d leads to an amazing 0.51*e* decrease in metal charge, clearly indicating its importance in the La(III) electronic structure. This result is not entirely unanticipated, as previous studies have noted the potential importance of the 5d orbitals, particularly as their energies are quite sensitive to the degree of relativistic effects.[6] Adding functions beyond the 6s6p5d4f has no substantial effect upon q_{La}, leading us to propose that these NAOs are the appropriate valence orbitals to calculate La(III) charges. This partitioning should also be appropriate for other Ln(III) cations, as going across the period constitutes filling the 4f shell.

Table 4. NPA Charges (q) Obtained with Different Density Functionals (at the Respective Optimized Geometries) Using Both the Modified (6s6p5d4f) and Default (6s4f for La(III) 6s5d for LuIII)) Valence Partition of the Small-Core RECP and the Segmented Contracted [9s8p5d4f3g] Basis

	modified valence		default valence	
	q_{La}	q_{Lu}	q_{La}	q_{Lu}
SVWN5	1.624	1.586	2.559	1.971
SPW92	1.623	1.586	2.558	1.971
PW91	1.821	1.766	2.622	2.107
PBE	1.863	1.771	2.650	2.111
TPSS	1.867	1.793	2.643	2.129
B3LYP	1.948	1.857	2.695	2.191
PBE0	1.925	1.867	2.689	2.205

In the case of $Lu(H_2O)_8{}^{3+}$, the default NPA partitioning places the 6s and 5d NAOs in the valence space, with the 4f in the core owing to the $4f^{14}$ electronic configuration. Similar results to those of $La(H_2O)_8{}^{3+}$ are obtained when the valence/Rydberg partitioning is modified; however, the relative contributions of the NAOs are somewhat altered (Table 3). As in $La(H_2O)_8{}^{3+}$, the metal charge is close to the formal charge of 3+ when all valence orbitals are placed in the Rydberg set. However, inclusion of the 6s in the valence partition decreases the atomic charge by 0.16e, nearly double that observed in La, and the 6p adds an additional 0.33e. The 5d adds 0.49e, which is slightly less than that found in $La(H_2O)_8{}^{3+}$. Including the 4f NAO in the valence space does not change q_{Lu} significantly, as it is doubly occupied and contributes exactly 2e when it is in the core space. The enhanced 6s and 6p contributions and decreased 5d participation in Lu(III) are in agreement with all electron DFT calculations of $La(H_2O)_{8,9}{}^{3+}$ and $Lu(H_2O)_{8,9}{}^{3+}$ using a zeroth-order regular approximation Hamiltonian, which indicates that the 5d orbitals are destabilized and the 6s orbitals are stabilized by increasing relativistic effects.[6] Since relativity plays an increasingly important role as one goes across the Ln period, the 6s orbital is more active in the metal−water interaction of $Lu(H_2O)_8{}^{3+}$ than in $La(H_2O)_8{}^{3+}$. However, the importance of the 6s is only manifested in the NPA and not within natural bond order analysis (NBO). Indeed, NBO predicts no covalent interaction between the Ln and ligating waters, irrespective of the partitioning of the valence space.

Using both the modified NPA and default partitioning schemes, the dependence of the metal charge upon basis set and density functional has also been examined. Table 4 shows the calculated charges on the metal center in $La(H_2O)_8{}^{3+}$ and $Lu(H_2O)_8{}^{3+}$ with different functionals at their respective optimized geometries. Both the modified and default NPA schemes predict that the metal charges increase as LSDA < GGA < meta-GGA < hybrid and that these charges span ~0.3e. Yet, the two partitions differ significantly when comparing q_{La} and q_{Lu}. Specifically, the default NPA charges predict that La(III) has nearly 0.5e less than Lu(III) in $Ln(H_2O)_8{}^{3+}$, while the modified NPA charges are within 0.1e for the two metals. This is an important observation if one is interested in calculating the relative surface charge density, ρ_s, of La(III) and Lu(III), as the default partition would significantly overemphasize the electrostatic differences across the Ln period.

Table 5. B3LYP NPA Charges (q) Using the Modified (6s6p5d4f) Valence Partition as a Function of Segmented Contracted Basis Set Using the Small-Core (ECP28MWB) RECP at the Optimized [9s8p5d4f3g] Geometry of $Ln(H_2O)_8{}^{3+}$ and Using the Large-Core (ECP47MWB in La and ECP60MWB in Lu) RECP at the Optimized Geometry

	q_{La}	q_{Lu}
	ECP28MWB	
[9s8p5d4f3g]	1.948	1.857
[9s8p5d3f]	2.035	1.900
[9s7p5d3f]	2.240	2.190
[9s6p5d3f]	2.442	2.315
[8s6p5d3f]	2.518	2.480
[7s6p5d3f]	2.573	2.552
[6s6p5d3f]	3.561	3.590
	ECP47MWB	ECP60MWB
[5s4p3d]	2.049	1.804

A comparison of the NPA charges as a function of basis set yields several important observations (Table 5). First, choice of the large-core or small-core RECP has little effect upon the calculated metal charge. Second, the calculated charges as a function of basis set truncation level qualitatively reflects the importance of key functions in the basis. For example, the g functions have neither energetic nor electrostatic importance, and truncation to the [6s6p5d3f] level yields completely unrealistic energies and charges for $Ln(H_2O)_8{}^{3+}$. In between these two extremes, NPA predicts roughly the same contributions of each function to the electrostatic description.

Conclusions

Benchmark calculations on $Ln(H_2O)_{1,8-9}{}^{3+}$ (Ln = La, Lu) have examined changes in predicted geometric and electronic structure using different density functionals and basis sets. Using the small-core RECP, we have highlighted specific functions for La^0 and Lu^0 that must be included in the basis for an adequate description of the Ln AOs. Using the (13s13p10d8f6g) uncontracted basis for La^{3+}, these are the 11th and 12th s functions and the ninth p function, while for Lu^{3+} the 10th through 12th s functions, the 12th p function, and the most diffuse f function are important. Differences in the calculated La^{3+} and Lu^{3+} HF and B3LYP cation densities have been identified as the source of a significant B3LYP contraction error in Lu^{3+}. As previously reported, large-core RECP calculations predict longer Ln−ligand bond lengths relative to small-core calculations. However, we have also shown that deviations in structural parameters extend beyond the Ln−ligand bonds and alter the H-bond distances in the primary hydration shell. In accordance with previous studies, we observe overbinding when using the LSDA functionals. The meta-GGA TPSS functional has the closest structural agreement of its optimized geometries relative to experimental results; however, the performance of the common B3LYP functional is also reasonable. Interestingly, LSDA, TPSS, and B3LYP calculate nearly identical lanthanide contraction values, while GGA, PBE, and PB0 have shorter values. In combination with our basis set results, this highlights the misleading nature of the lanthanide contraction as a metric for assessing the quality of method/basis set combinations. In fact, Δ_{Ln} is remarkably insensitive to clear inadequacies of a basis due to cancellation of errors. Finally, we have examined the calculated charges from natural population analyses and proposed that the 6s6p5d4f NAOs be placed in the valence space when determining atomic charge. Clear deviations in the electrostatic description are observed when the NPA charges from the modified and default partitioning schemes are compared.

Acknowledgment. This work was sponsored by the U.S. Department of Energy, Office of Nuclear Energy, Science and Technology, Junior Faculty Award Program award #DE-FG07-05ID14692/IDNE006. This research was performed in part using (1) the Molecular Science Computing Facility (MSCF) in the William R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory, operated for the Department of Energy by Battelle, and (2) the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science user facility at Lawrence Berkely National Laboratory. The author would also like to thank Prof. Kirk Peterson for several thoughtful discussions.

References

(1) Nielson, G. W.; Skipper, N. T. K^+ Coordination in Aqueous Solution. *Chem. Phys. Lett.* **1985**, *114*, 35–38.

(2) Neilson, G. W.; Broadbent, R. D. The Structure of Sr^{2+} in Solution. *Chem. Phys. Lett.* **1990**, *167*, 429–431.

(3) Salmon, P. S.; Neilson, G. W.; Enderby, J. E. The of Cu^{2+} Aqueous Solutions. *J. Phys. C.: Solid State Phys.* **1988**, *21*, 1335–1350.

(4) Choppin, G. R.; Rizkalla, E. N. Solution Chemistry of Actinides and Lanthanides. *Handb. Phys. Chem. Rare Earths* **1994**, *18*, 559–590.

(5) Kuchle, W.; Dolg, M.; Stoll, H. Ab Initio Study of the Lanthanide and Actinide Contraction. *J. Phys. Chem. A* **1997**, *101*, 7128–7133.

(6) Clavaguera, C.; Gognon, J. P.; Pyykko, P. Calculated Lanthanide Contractions for Molecular Trihalides and Fully Hydrated Ions: The Contributions from Relativity and 4f-shell Hybridization. *Chem. Phys. Lett.* **2006**, *429*, 8–12.

(7) Laerdahl, J. K.; Faegri, K., Jr.; Visscher, L.; Saue, T. A fully relativistic Dirac–Hartree–Fock and second-order Møller–Plesset study of the lanthanide and actinide contraction. *J. Chem. Phys.* **1998**, *109*, 10806–10816.

(8) Pyykko, P. Relativistic Effects in Structural Chemistry. *Chem. Rev.* **1988**, *88*, 563–594.

(9) Maron, L.; Eisenstein, O. Do f Electrons Play a Role in the Lanthanide-Ligand Bonds? A DFT Study of $Ln(NR_2)_3$; R = H, SiH_3. *J. Phys. Chem. A* **2000**, *104*, 7140–7143.

(10) Clark, D. L.; Gordon, J. C.; Hay, P. J.; Martin, R. L.; Poli, R. DFT Study of Tris(bis(trimethylsilyl)methyl)lanthanum and -samarium. *Organometallics* **2002**, *21*, 5000–5006.

(11) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural Population Analysis. *J. Chem. Phys.* **1985**, *83*, 735–746.

(12) Slater, J. C. *Quantum Theory of Molecular and Solids*; McGraw-Hill: New York, 1974; Vol. 4: The Self-Consistent Field for Molecular and Solids.

(13) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

(14) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(15) Vosko, S. H.; Wilke, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(16) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244–13249.

(17) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(18) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(19) Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.

(20) Perdew, J. P. *Electronic Structure of Solids*; Ziesche, P., Eschrig, H., Eds.; Akademie: Berlin, 1991.

(21) Burke, K.; Perdew, J. P.; Wang, Y. *Electronic Density Functional Theory: Recent Progress and New Directions*; Dobson, J. F., Vignale, G., Das, M. P., Eds.; Plenum Press: New York, 1998.

(22) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671–6687; **1993**, *48*, 4978(E).

(23) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(24) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.

(25) (a) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627. (b) Hertwig, R. H.; Koch, W. On the Parameterization of the Local Correlation Functional. What is Becke-3-LYP? *Chem. Phys. Lett.* **1997**, *268*, 345.

(26) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof Exchange-Correlation Functional. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

(27) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solid. *Phys. Rev. Lett.* **2003**, *91*, 146401–146405.

(28) Bylaska, E. J.; de Jong, W. A.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Aprà, E.; Windus, T. L.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.;

Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, Version 5.0; Pacific Northwest National Laboratory: Richland, WA, 2006.

(29) 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.

(30) Cao, X.; Dolg, M. Valence Basis Sets for Relativistic Energy-Consistent Small-Core Lanthanide Pseudopotentials. *J. Chem. Phys.* **2001**, *115*, 7348–7355.

(31) Cao, X.; Dolg, M. Segmented Contraction Scheme for Small-Core Lanthanide Pseudopotential Basis Sets. *THEOCHEM* **2002**, *581*, 139.

(32) Dolg, M.; Stoll, H.; Savin, A.; Preuss, H. Pseudopotential Study of the Rare Earth Monohydrides, Monoxides, and Monofluorides. *Theor. Chim. Acta* **1989**, *75*, 173–194.

(33) (a) Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023. (b) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(34) Mulliken, R. S. Electronic Population Analysis on LCAOMO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

(35) Shannon, R. D. Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides. *Acta Crystallogr., Sect. A* **1976**, *32*, 751–767.

(36) Quadrelli, E. A. Lanthanide Contraction over the 4f Serie Followsa Quadratic Decay. *Inorg. Chem.* **2002**, *41*, 167–169.

(37) Seitz, M.; Oliver, A. G.; Raymond, K. N. The Lanthanide Contraction Revisited. *J. Am. Chem. Soc.* **2007**, *129*, 11153–11160.

(38) Deng, B.; Ellis, D. E.; Ibers, J. A. New Layered Rubidium Rare-Earth Selenides: Syntheses, Structures, Physical Properties, and Electronic Structures for $RbLnSe_2$. *Inorg. Chem.* **2002**, *41*, 5716–5720.

(39) Yao, J.; Deng, B.; Sherry, L. J.; McFarland, A. D.; Ellis, D. E.; Van Duyne, R. P.; Ibers, J. A. Syntheses, Structure, Some Band Gaps, and Electronic Structures of $CsLnZnTe_3$ (Ln = La, Pr, Nd, Sm, Gd, Tb, Dy, Ho, Er, Tm, Y). *Inorg. Chem.* **2004**, *43*, 7735–7740.

(40) Baisch, U.; Belli Dell'Amico, D.; Calderazzo, F.; Conti, R.; Labella, L.; Marchetti, F.; Quadrelli, E. A. The Mononuclear and Dinuclear Dimethoxyethane Adducts of Lanthanide Trichlorides $[LnCl_3(DME)_2)_n$, $n = 1$ or 2, Fundamental Starting Materials in Lanthanide Chemistry: Preparations and Structures. *Inorg. Chim. Acta* **2004**, *357*, 1538–1548.

(41) Baisch, U.; Belli Dell'Amico, D.; Calderazzo, F.; Labella, L.; Marchetti, F.; Merigo, A. N,N-Dialkylcarbamato Lanthanide Complexes, a series of Isotypical Coordination Compounds. *Eur. J. Inorg. Chem.* **2004**, 1219.

(42) Slater, J. C. Atomic Shielding Constants. *Phys. Rev.* **1930**, *36*, 57–64.

(43) Clementi, E.; Raimondi, D. L. Atomic Screening Constants from SCF Functions. *J. Chem. Phys.* **1963**, *38*, 2686–2689.

(44) Clementi, E.; Raimondi, D. L.; Reinhardt, W. P. Atomic Screening Constants from SCF Functions. II. Atoms with 37 to 86 Electrons. *J. Chem. Phys.* **1967**, *47*, 1300–1307.

(45) Pyykko, P. Dirac-Fock One Centre Calculations Part 8. The 1Σ States of ScH, YH, LaH, AcH, TmH, LuH and LrH. *Phys. Scr.* **1979**, *20*, 647–651.

(46) Schwarz, W. H. E. Relativistic calculations of molecules relativity and bond lengths. *Phys. Scr.* **1987**, *36*, 403–411.

(47) Habenschuss, A.; Spedding, F. H. The coordination (hydration) of rare earth ions in aqueous chloride solutions from X-Ray diffraction. I. $TbCl_3$, $DyCl_3$, $ErCl_3$, $TmCl_3$, and $LuCl_3$. *J. Chem. Phys.* **1979**, *70*, 2797–2806.

(48) Habenschuss, A.; Spedding, F. H. The coordination (hydration) of rare earth ions in aqueous chloride solutions from x-ray diffraction. II. $LaCl_3$, $PrCl_3$, and $NdCl_3$. *J. Chem. Phys.* **1979**, *70*, 3758–3763.

(49) Yamauchi, S.; Kanno, H.; Akama, Y. Far-Infrared Evidence for the Hydration Number Change of Rare Earth Ions in Aqueous Solution. *Chem. Phys. Lett.* **1988**, *151*, 315–317.

(50) Mochizuki, Y.; Tatewaki, H. *Chem. Phys.* **2001**, *273*, 135–148.

(51) Clark, A. E.; Martin, R. L.; Hay, P. J.; Green, J. C.; Jantunen, K. C.; Kiplinger, J. L. Electronic structure, excited states, and photoelectron spectra of uranium, thorium, and zirconium bis(ketimido) complexes $(C_5R_5)_2$-$M[-N{=}CPh_2]_2$ (M = Th, U, Zr; R = H, CH_3). *J. Phys. Chem. A* **2005**, *109*, 5481–5491.

(52) Peralta, J. E.; Batista, E. R.; Scuseria, G. E.; Martin, R. L. All-Electron Hybrid Density Functional Calculations on UF_n and UCl_n ($n = 1$–6). *J. Chem. Theory Comput.* **2005**, *1*, 612–616.

(53) Perdew, J. P.; Schmidt, K. *Density Functional Theory and Its Application to Materials*; Van Doren, E., Ed.; AIP Press: Melville, NY, 2001.

(54) Vetere, V.; Maldivi, P.; Adamo, C. Comparative studies of quasi-relativistic density functional methods for the description

of lanthanide and actinide complexes. *J. Comput. Chem.* **2003**, *24*, 850.

(55) Adamo, C.; Maldivi, P. A Theoretical Study of Bonding in Lanthanide Trihalides by Density Functional Methods. *J. Phys. Chem. A* **1998**, *102*, 6812–6820.

(56) Petit, L.; Borel, A.; Daul, C.; Maldivi, P.; Adamo, C. A Theoretical Characterization of Covalency in Rare Earth Complexes through Their Absorption Electronic Properties: f-f Transitions. *Inorg. Chem.* **2006**, *45*, 7382–7388.

(57) Gutowski, K. E.; Dixon, D. A. Predicting the Energy of the Water Exchange Reaction and Free Energy of Solvation for the Uranyl Ion in Aqueous Solution. *J. Phys. Chem. A* **2006**, *110*, 8840–8856.

(58) Cundari, T. R.; Stevens, W. J. Effective Core Potential Methods for Lanthanides. *J. Chem. Phys.* **1993**, *98*, 5555–5565.

(59) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. Relativistic Compact Effective Potentials and Efficient, Shared-Exponent Basis Sets for the Third-, Fourth-, and Fifth-Row Atoms. *Can. J. Chem.* **1992**, *70*, 612–630.

(60) Dinescu, A.; Clark, A. E. Thermodynamic and Structural Features of Aqueous Ce(III). *J. Phys. Chem. A*, in preparation.

(61) (a) Löwdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J. Chem. Phys.* **1950**, *18*, 365. (b) Löwdin, P.-O. On the Non-Orthogonality Problem. *Adv. Quantum Chem.* **1970**, *5*, 185.

(62) Maseras, F.; Morokuma, K. Application of the natural population analysis to transition-metal complexes. Should the empty metal p orbitals be included in the valence space. *Chem. Phys. Lett.* **1992**, *195*, 500–504.

(63) Clark, A. E.; Sonnennberg, J.; Hay, P. J.; Martin, R. L. Density and wave function analysis of actinide complexes: What can fuzzy atom, atoms-in-molecules, Mulliken, Löwdin, and natural population analysis tell us. *J. Chem. Phys.* **2004**, *121*, 2563–2570.

(64) Dobler, M.; Guilbaud, P.; Dedieu, A.; Wipff, G. Interaction of Trivalent Lanthanide Cations with Nitrate Anions: A Quantum Chemical Investigation of Monodentate/Bidentate Binding Modes. *New J. Chem.* **2001**, *25*, 1458–1465.

(65) Saloni, J.; Roszak, S.; Hilpert, K.; Popovic, A.; Miller, M.; Leszczynski, J. Mass Spectrometric and Quantum Chemical Studies of the Thermodynamics and Bonding of Neutral and Ionized LnCl, $LnCl_2$, and $LnCl_3$ Species (Ln = Ce, Lu). *Inorg. Chem.* **2006**, *45*, 4508–4517.

CT700317P

JCTC Journal of Chemical Theory and Computation

Basis Set Convergence of Nuclear Magnetic Shielding Constants Calculated by Density Functional Methods

Frank Jensen

*Department of Chemistry, University of Aarhus, Langelandsgade 140,
DK-8000 Aarhus, Denmark*

Received January 11, 2008

Abstract: The previously proposed polarization consistent basis sets, optimized for density functional calculations, are evaluated for calculating nuclear magnetic shielding constants. It is shown that the basis set convergence can be improved by adding a single p-type function with a large exponent and allowing for a slight decontraction of the p functions. The resulting pcS-*n* basis sets should be suitable for calculating nuclear magnetic shielding constants with density functional methods and are shown to perform significantly better than existing alternatives for a comparable computational cost.

I. Introduction

The use of nuclear magnetic resonance methods for probing molecular structures in solution-phase environments is well established, and technical and methodology improvements continue to push the limits for the size of molecules that can be handled. An increasingly important element for interpreting the experimental data is the simultaneous calculation of spectral information, which allows a direct correlation between molecular structure and quantities such as nuclear magnetic shielding and spin−spin coupling constants.[1] In order for this to become a routine procedure, it is necessary that there exist computational procedures that both are fast and have well-defined accuracies. While sophisticated methods such as coupled cluster can provide very accurate results, they are limited to relatively small systems.[2] Density functional methods,[3] on the other hand, are applicable to systems with hundreds or even thousands of atoms.[4] The main drawback of density functional methods is the inability to systematically improve the results, but Keal and Tozer have recently proposed new exchange-correlation functionals aimed at calculating nuclear magnetic shielding constants.[5]

The second component in performing electronic structure calculations is the use of a basis set for expanding the molecular orbitals. A large basis set will enable the full potential of the chosen method for calculating the wave function to be realized but also requires a large computational cost. A small basis set, on the other hand, is computationally efficient but introduces errors in the results. It is therefore desirable to have a sequence of basis sets such that the accuracy can be controlled and assessed at each level, and at the same time being as compact as possible. For wave function based methods including electron correlation, the correlation consistent (cc-pVXZ)[6] basis sets developed by Dunning and co-workers represent such a hierarchy for energies and structural properties. For independent particle models, such as density functional theory, we have developed the polarization consistent (pc-*n*)[7] basis sets for providing a fast and controlled convergence toward the basis set limit.

In discussions of basis set convergence, it is important to realize that different methods and properties have different basis set requirements and convergence rates. Both the cc-pVXZ and pc-*n* basis sets have been constructed using energetic criteria, such that functions which contribute similar amounts of energy are included at the same stage, and this leads to the maximum angular momentum function included in the basis set as the natural expansion parameter. The differences between the cc-pVXZ and pc-*n* basis sets are related to the fact that the electron correlation energy converges as an inverse polynomial in the maximum angular momentum function,[8] while the density functional energy displays an exponential convergence.[9] By virtue of construction, the cc-pVXZ and pc-*n* basis sets are therefore expected to provide the fastest convergence toward the limiting value for correlation and density functional energies, respectively. Energetically related properties, such as equilibrium geometries and vibrational frequencies, usually also display a smooth convergence toward the basis set limit.[10]

10.1021/ct800013z CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/05/2008

For molecular properties that depend on the energetically unimportant region far from the nuclei, such as electric multipole moments and polarizabilities, the basis set convergence can be substantially improved by adding diffuse functions, leading to the aug-cc-pVXZ[11] and aug-pc-*n*[12] basis sets. In recent work we have shown how the basis set convergence for calculating nuclear spin−spin coupling constants can be improved by adding tight functions, leading to the definition of the pcJ-*n* basis sets.[13] The spin−spin coupling constant has four independent contributions, and an interesting observation was that the different operators have different basis set requirements. The Fermi-contact operator is only sensitive to the presence of (tight) s-type functions and the paramagnetic spin−orbit (PSO) operator is only sensitive to the presence of p-type functions, while the spin-dipole operator is sensitive to p-, d-, and f-type functions. In order to ensure a fast basis set convergence of the spin−spin coupling constant, it was therefore necessary to add tight s-, p-, d-, and f-type basis functions.

The nuclear magnetic shielding constant $\boldsymbol{\sigma}$ can be defined as the second derivative of the energy with respect to an external magnetic field $\mathbf{B}$ and a nuclear magnetic moment $\mathbf{I}$. In a perturbation formulation, the shielding constant can be written in terms of a diamagnetic and paramagnetic contribution, where the former is calculated as an expectation value of the diamagnetic shielding operator ($\mathbf{H}^{\mathrm{DS}}$) while the latter is calculated as a response property of the paramagnetic spin−orbit ($\mathbf{H}^{\mathrm{PSO}}$) and orbital Zeeman ($\mathbf{L}_G$) operators.[14,15]

$$\begin{aligned}\sigma &= \frac{\partial^2 E}{\partial \mathbf{B}\,\partial \mathbf{I}} \\ &= \langle \Psi_0 | \mathbf{H}^{\mathrm{DS}} | \Psi_0 \rangle - 2\sum_{n\neq 0} \frac{\langle \Psi_0 | \mathbf{H}^{\mathrm{PSO}} | \Psi_n \rangle \langle \Psi_n | \mathbf{L}_G | \Psi_0 \rangle}{E_0 - E_n} \\ \mathbf{H}^{\mathrm{DS}} &= \frac{g_A \mu_N}{2c^2} \frac{\mathbf{r}_{iG}^t \mathbf{r}_{iA} - \mathbf{r}_{iA} \mathbf{r}_{iG}^t}{r_{iA}^3} \\ \mathbf{H}^{\mathrm{PSO}} &= \frac{g_A \mu_N}{c^2} \frac{\mathbf{r}_{iA} \times \mathbf{p}_i}{r_{iA}^3} \\ \mathbf{L}_G &= \frac{1}{2} \mathbf{r}_{iG} \times \mathbf{p}_i \end{aligned} \tag{1}$$

Here $\mathbf{r}_{iA/G}$ denoted the position vector between electron i and nucleus A or the gauge origin G, μ_N is the nuclear magneton, and g_A is the nuclear g factor. The nuclear magnetic shielding constant is a 3 × 3 tensor, but only the average isotropic component corresponding to one-third of the trace of $\boldsymbol{\sigma}$ is observed in solution, and we will consequently focus on this. It is customary to use the units of ppm, and this will also be the case here.

Given our findings for the basis set requirements of the PSO operator, it follows that the basis set convergence for nuclear magnetic shielding constants potentially could be improved by adding tight p functions. Furthermore, to our knowledge detailed basis set requirements of the diamagnetic shielding and orbital Zeeman terms have not been investigated. The present paper examines the basis set convergence of the nuclear magnetic shielding constant using density functional methods in more detail.

A large body of previous work on calculating nuclear magnetic shielding constants exists, with the majority employing standard basis sets, such as the Pople style *k-lm*G basis sets,[16] the Ahlrichs basis sets of double, triple, and quadruple quality,[17] and the cc-pVXZ[18] and pc-*n*[19] families of basis sets. The IGLO basis sets have been proposed explicitly for magnetic properties,[20] and Manninen and Vaara have proposed to use basis sets complete to within a given threshold in a given exponent range for calculating magnetic properties, but only a single basis set suitable for first-row elements and hydrogen has been defined.[21]

II. Computational Details

All calculations have been performed with the Dalton[22] and Gaussian-03[23] program packages using the KT3 and B3LYP[24] exchange-correlation functionals. The GIAO technique has been employed to ensure gauge independence of the calculated results.[25] Molecular geometries have been taken from the G3 data set[26] or optimized at the B3LYP/6-31G(d,p) level. We emphasize that only the convergence with respect to the basis set is investigated in the present case, and no attempt is made to compare with experimental results. The latter requires attention to the molecular geometry and the quality of the exchange-correlation functional as well as vibrational and environmental effects. It is demonstrated that the basis set convergence is very similar for the two employed functionals, and the basis set convergence should therefore be representative for Hartree−Fock and density functional methods in general.

III. Improved Basis Sets for Calculating Nuclear Magnetic Shielding Constants

The notation for the polarization consistent basis sets is pc-*n*, where *n* indicates the level of polarization beyond the atomic system, i.e. pc-0 is unpolarized, pc-1 is of double-ζ quality with a single polarization function, pc-2 is of triple-ζ quality with d- and f-type polarization functions, etc. An initial exploration using the uncontracted pc-*n* basis sets showed that only p-type tight functions had any significant influence on the calculated nuclear magnetic shielding constants, which is consistent with the findings for the PSO operator in the previous study. Diffuse functions in some cases also had a significant effect, which may be related to the orbital Zeeman operator or simply to the fact that polar systems with lone pairs in general require diffuse functions for an adequate description. The diamagnetic shielding operator was not found to have additional basis set requirements beyond those already included for representing the electron density.

In order to determine the optimum exponents for the tight p functions, we employed an optimization procedure analogous to that used for defining the pcJ basis sets where the optimization criterion is to maximize the change in the nuclear shielding constant relative to the regular pc-*n* basis set. The optimum exponents determined for a selection of molecular systems showed a high degree of regularity with a near-constant ratio of 6.5 relative to the highest exponent already included in the basis set. Addition of a second tight

Table 1. Basis Set Composition in Terms of Uncontracted Functions

basis	H	Li−Be	B−Ne	Na−Mg	Al−Ar
pc-0	3s	5s1p	5s3p	8s5p	8s6p
pc-1	4s1p	7s3p	7s4p1d	11s7p	11s8p1d
pc-2	6s2p1d	10s4p1d	10s6p2d1f	13s9p1d	13s10p2d1f
pc-3	9s4p2d1f	14s6p2d1f	14s9p4d2f1g	17s12p2d1f	17s13p4d2f1g
pc-4	11s6p3d2f1g	19s8p3d2f1g	18s11p6d3f2g1h	21s15p3d2f1g	21s16p6d3f2g1h
pcS-0	3s	5s2p	5s4p	8s6p	8s7p
pcS-1	4s2p	7s4p	7s5p1d	11s8p	11s9p1d
pcS-2	6s3p1d	10s5p1d	10s7p2d1f	13s10p1d	13s11p2d1f
pcS-3	9s5p2d1f	14s7p2d1f	14s10p4d2f1g	17s13p2d1f	17s14p4d2f1g
pcS-4	11s7p3d2f1g	19s9p3d2f1g	18s12p6d3f2g1h	21s16p3d2f1g	21s17p6d3f2g1h
STO-3G	3s	6s3p	6s3p	9s6p	9s6p
6-31G(d,p)	4s1p	10s4p1d	10s4p1d	16s10p1d	16s10p1d
6-311G(2df,2pd)[a]	5s2p1d	11s5p2d1f	11s5p2d1f	13s9p2d1f	13s9p2d1f
cc-pVDZ	4s1p	9s4p1d	9s4p1d	12s8p1d	12s8p2d
cc-pVTZ	5s2p1d	11s5p2d1f	10s5p2d1f	15s10p2d1f[f]	15s9p3d1f
cc-pVQZ	6s3p2d1f	12s6p3d2f1g	12s6p3d2f1g	16s12p3d2f1g[g]	16s11p4d2f1g
cc-pV5Z	8s4p3d2f1g	14s8p4d3f2g1h[b]	14s8p4d3f2g1h	20s14p4d3f2g1h[h]	20s12p5d3f2g1h
cc-pCVDZ	4s1p	9s4p1d	10s5p1d	13s9p2d	13s9p2d
cc-pCVTZ	5s2p1d	11s5p2d1f	12s7p3d1f	17s12p4d2f[i]	17s11p4d2f
cc-pCVQZ	6s3p2d1f	12s6p3d2f1g	15s9p5d3f1g	19s15p6d4f2g[j]	19s14p6d4f2g
SVP	4s1p	7s4p[c]	7s4p1d	10s6p	10s7p1d
TZV	5s2p1d	11s4p[d]	11s6p2d1f	14s8p1d	14s9p2d1f
QZV	7s3p2d1f	15s7p2d1f[e]	15s7p2d1f	20s12p3d1f	20s14p4d2f1g
IGLO-II	5s1p		9s5p1d		11s7p2d
IGLO-III	6s2p		11s7p2d		12s8p3d

[a] McLean-Chandler basis set[30] for Na−Ar. [b] 7p for Li. [c] 1p for Li. [d] 3p for Li. [e] 6p for Li. [f] 16s for Na. [g] 19s for Na. [h] 19s12p for Na. [i] 18s for Na. [j] 22s for Na.

p function was in all cases found to give almost negligible changes. These findings are in complete agreement with the previous study for the PSO operator and suggest that a faster basis set convergence can be obtained by adding a single tight p-type function to the regular pc-*n* basis sets.

The lighter s-block elements (H, He, Li, Be) do not have occupied p orbitals, and it is therefore not a priori clear whether the results for these elements will be sensitive to the presence of tight p functions. In test calculations we found that the results for these elements are less affected by tight p functions than for the p-block elements but that a systematic improvement does occur, and we have therefore elected to include a tight p function also for these elements. We thus define a sequence of pc-type basis sets optimized for calculating nuclear magnetic shielding constants by adding a single tight p-type function to the regular pc-*n* basis sets to produce a basis set denoted pcS-*n*, where S indicates shielding. The size of the basis sets are shown in Table 1.

The pc-*n* basis sets employ a general contraction scheme using atomic orbital coefficients, where the degree of contraction is determined by the criterion that the contraction error should be smaller than the error of the uncontracted basis set relative to the basis set limit. For the pc-*n* basis sets this leads automatically to the pc-1 basis set being of double-ζ quality in the valence region, the pc-2 basis set being of triple-ζ quality, etc. When this criterion is used for the pcJ-*n* basis sets, it allows only a small degree of contraction, as the nuclear spin−spin coupling constant requires substantial flexibility in the core region. We have employed the same strategy in the present case, and it is found that the nuclear magnetic shielding constant requires more flexibility in the inner valence region for the p orbitals than the regular pc-*n* basis sets, and the recommended contractions are shown in Table 2. The pcS-0 basis set is of the same size as the pc-0 basis set, while the pcS-1 and pcS-2 basis sets have one or two more (contracted) p functions. The pcS-3 basis set requires further decontraction of the p functions to ensure that the contraction error remains below the inherent error of the uncontracted basis set. The pcS-4 basis set is sufficiently large that the remaining basis set errors are only fractions of a ppm, and it is difficult to devise a contraction scheme without degrading this accuracy. We have chosen the contraction shown in Table 2, where the s contraction is relaxed for second-row elements relative to the pc-4 basis set and the p functions are almost uncontracted.

IV. Basis Set Convergence

We have examined the performance of the pcS-*n* and aug-pcS-*n* families of basis sets, as well as a selection of other popular basis sets, shown in Tables 1 and 2, for the systems in Table 3. As there are some differences between the nuclear magnetic shielding constants, the results will be divided into five groups: Hydrogen (H), first- (M_1 = Li, Be) and second-row (M_2 = Na, Mg) metallic elements, and first- (A_1 = C−F) and second -row (A_2 = Si−Cl) main-group elements. The basis set limiting value is in each case taken as the uncontracted aug-pcS-4 value, except for the largest systems where the pcS-4 results were used. The difference in the results with the (uncontracted) aug-pcS-3 and aug-pcS-4 basis sets indicates that the reference values are converged well below 0.001 ppm for hydrogen and 0.1 ppm for the remaining elements. A couple of pathological cases where this is not the case are discussed at the end of this section.

Table 4 shows the (uncontracted) basis set errors relative to the aug-pcS-4 results quantified in terms of the mean absolute deviation (MAD) over the symmetry-unique shielding constants using the KT3 and B3LYP exchange-correla-

Table 2. Basis Set Composition in Terms of Contracted Functions

basis	H	Li−Be	B−Ne	Na−Mg	Al−Ar
pc-0	2s	3s1p	3s2p	4s2p	4s3p
pc-1	2s1p	3s2p	3s2p1d	4s2p	4s3p1d
pc-2	3s2p1d	4s2p1d	4s3p2d1f	5s3p1d	5s4p2d1f
pc-3	5s4p2d1f	6s3p2d1f	6s5p4d2fg1	6s4p2d1f	6s5p4d2f1g
pc-4	7s6p3d2f1g	8s4p3d2f1g	8s7p6d3f2g1h	7s5p3d2f1g	7s6p6d3f2g1h
pcS-0	2s	3s1p	3s2p	4s2p	4s3p
pcS-1	2s1p	3s3p	3s3p1d	4s4p	4s4p1d
pcS-2	3s2p1d	4s3p1d	4s4p2d1f	5s6p1d	5s6p2d1f
pcS-3	5s4p2d1f	6s6p2d1f	6s8p4d2fg1	7s8p2d1f	7s9p4d2f1g
pcS-4	7s6p3d2f1g	8s8p3d2f1g	8s10p6d3f2g1h	10s11p3d2f1g	10s12p6d3f2g1h
STO-3G	1s	2s1p	2s1p	3s2p	3s2p
6-31G(d,p)	2s1p	3s2p1d	3s2p1d	4s3p1d	4s3p1d
6-311G(2df,2pd)[a]	3s2p1d	4s3p2d1f	4s3p2d1f	6s5p2d1f	6s5p2d1f
cc-pVDZ	2s1p	3s2p1d	3s2p1d	4s3p1d	4s3p2d
cc-pVTZ	3s2p1d	4s3p2d1f	4s3p2d1f	5s4p2d1f	5s4p3d1f
cc-pVQZ	4s3p2d1f	5s4p3d2f1g	5s4p3d2f1g	6s5p3d2f1g	6s5p4d2f1g
cc-pV5Z	5s4p3d2f1g	6s5p4d3f2g1h	6s5p4d3f2g1h	7s6p4d3f2g1h	7s6p5d3f2g1h
cc-pCVDZ	2s1p	4s3p1d	4s3p1d	5s4p2d	5s4p2d
cc-pCVTZ	3s2p1d	6s5p3d1f	6s5p3d1f	7s6p4d2f	7s6p4d2f
cc-pCVQZ	4s3p2d1f	8s7p5d3f1g	8s7p5d3f1g	9s8p6d4f2g	9s8p6d4f2g
SVP	2s1p	3s2p[b]	3s2p1d	4s2p	4s3p1d
TZV	3s2p1d	5s3p[c]	5s3p2d1f	5s4p1d	5s4p2d1f
QZV	4s3p2d1f	7s4p2d1f[d]	7s4p3d2f1g	9s5p4d1f[e]	9s6p4d2f1g
IGLO-II	3s1p		5s4p1d		7s6p2d
IGLO-III	4s2p		7s6p2d		8s7p3d

[a] McLean-Chandler basis set[30] for Na−Ar. [b] 1p for Li. [c] 2p for Li. [d] 6s for Li. [e] 3d for Na.

Table 3. Molecular Systems Used for Calibration

CH_4, NH_3, H_2O, HF, N_2, F_2, CO, CO_2, F_2O
C_2H_2, C_2H_4, C_2H_6, H_2CO, HCOOH, H_2CS, N_2H_2, N_2H_4
CH_3NH_2, CH_3NO_2, CH_3OH, CH_3F, CH_3CN, CH_2F_2, CH_3CHO, H_2CCHCN
CH_3SiH_3, CH_3PH_2, CH_3SH, CH_3Cl, CH_2Cl_2, C_2N_2, C_2F_4, C_2Cl_4
cyclopropene, butadiene, benzene, furan, pyrrole, thiophene, pyridine, $(CH_3)_2SO$
LiH, $LiCH_3$, Li_2O, LiF, Li_2S_2, LiCl
Be_2H_4, $Be(CH_3)_2$, BeF_2, $BeCl_2$
SiH_4, PH_3, H_2S, HCl, Si_2H_2
P_2, Cl_2, CS, CS_2, CSO, PF_5, PCl_5, SF_6, Cl_2SO_2
NaH, $NaCH_3$, Na_2O, NaF, Na_2S, NaCl
Mg_2H_4, $Mg(CH_3)_2$, MgF_2, $MgCl_2$

tion functionals. It is seen that addition of both diffuse functions (aug-pc-*n*) and a tight p function (pcS-*n*) to the pc-*n* basis sets has an effect and, with a few exceptions, leads to lower basis set errors. It is noticeable that in some cases there is a cooperative effect, where addition of both tight and diffuse functions leads to a larger improvement than the sum of the two individual effects. At the pc-1 level with the KT3 functional for first-row nonmetallic elements (A_1 = C−F), for example, the addition of diffuse functions improves the MAD by 2.7 ppm, the addition of a tight p function lowers the MAD by 10.1 ppm, but the combined effect is 15.2 ppm. These values should be compared to the MAD value of 18.0 ppm for the pc-1 basis set, which consequently is lowered to 2.8 ppm with the aug-pcS-1 basis set.

The basis set errors are much smaller for hydrogen shieldings than for the other elements, and the metallic s-block elements (Li, Be, Na, Mg) have lower errors at a given level than the p-block elements (C−F, Si−Cl). This is consistent with the importance of p functions for the PSO operator and the dominance of s-type bonding for hydrogen and the metallic elements. The basis set error for the second-row elements Si−Cl tend to be somewhat larger than for the first-row elements C−F, and the improvement by adding diffuse and tight functions is smaller. Table 4 shows that addition of a single tight p function can improve the basis set convergence, although the improvement is not as spectacular as for the spin−spin coupling constants. It can also be noted that the basis set convergence is very similar for the two employed exchange-correlation functionals.

Table 5 compares the performance of the (contracted) pcS-*n* and aug-pcS-*n* basis sets with a selection of other commonly used basis sets for the set of systems in Table 3. The Pople type STO-3G,[27] 6-31G(d,p),[28] and 6-311G-(2df,2pd)[29,30] basis sets represent minimum and double- and triple-ζ quality, and the last two can be augmented with diffuse functions (6-31++G(d,p) and 6-311++G(2df,2pd)). The Dunning family of correlation consistent basis sets cc-pVXZ (X = D, T, Q, 5)[31] can be augmented with both diffuse (aug-cc-pVXZ)[32] and tight functions (cc-pCVXZ).[33] The latter has been designed for recovering core and core-valence correlation and adds tight functions of all types as well as multiple sets of tight functions for the T and Q basis sets (Table 2). The Ahlrichs-type basis sets SVP,[34] TZP,[35] and QZP[36] are of double-, triple-, and quadruple-ζ quality, but no standard sets of diffuse and tight functions have been defined. The IGLO basis sets have been designed for magnetic properties but are not defined for s-block elements. They are furthermore somewhat difficult to classify in terms of quality, as the number of s and p functions (Table 2) indicates at least quadruple-ζ quality, but the lack of high angular momentum functions suggests that they are at best of triple-ζ quality. A comparison of the results in Table 5 for the KT3 and B3LYP functionals indicates a very similar basis set behavior, indicating that the conclusions discussed in the following should be valid for density functional methods in general.

Table 4. Mean Absolute Deviations (ppm) Relative to the Basis Set Limit for the Symmetry-Unique Nuclear Magnetic Shielding Constants for the Systems in Table *3*[a]

	KT3					B3LYP				
basis set	*H*	M_1	A_1	M_2	A_2	*H*	M_1	A_1	M_2	A_2
pc-0	1.5	14.9	63.5	7.4	81.0	1.5	15.1	64.2	8.8	90.6
aug-pc-0	1.2	12.9	32.8	7.9	54.0	1.1	13.8	30.8	8.6	60.5
pcS-0	1.5	9.3	67.0	6.8	88.9	1.5	8.1	65.0	8.0	99.8
aug-pcS-0	1.1	2.8	12.5	3.5	61.2	1.1	3.0	14.5	3.8	66.7
pc-1	0.38	2.8	18.0	2.7	18.3	0.37	2.8	17.8	2.9	19.7
aug-pc-1	0.25	2.9	15.3	2.5	16.8	0.25	2.9	15.9	2.6	19.3
pcS-1	0.25	1.1	7.9	2.6	19.1	0.24	1.2	7.7	2.8	20.4
aug-pcS-1	0.11	1.0	2.8	1.8	15.4	0.11	1.1	3.3	1.8	17.6
pc-2	0.14	0.94	4.27	0.95	6.65	0.14	0.94	3.92	1.05	7.19
aug-pc-2	0.10	0.87	2.65	1.17	3.96	0.10	0.95	2.83	1.17	3.88
pcS-2	0.073	0.50	2.88	0.75	6.50	0.074	0.45	2.28	0.84	6.92
aug-pcS-2	0.032	0.39	0.61	0.81	3.65	0.034	0.42	0.55	0.79	3.59
pc-3	0.016	0.16	0.49	0.19	0.95	0.017	0.16	0.40	0.19	0.85
aug-pc-3	0.013	0.16	0.25	0.19	0.44	0.014	0.16	0.29	0.18	0.43
pcS-3	0.008	0.07	0.32	0.18	0.96	0.008	0.07	0.23	0.18	0.84
aug-pcS-3	0.005	0.06	0.05	0.16	0.41	0.004	0.07	0.05	0.15	0.39
pc-4	0.001	0.052	0.108	0.023	0.118	0.001	0.048	0.087	0.017	0.110
pcS-4	0.001	0.005	0.059	0.020	0.121	0.000	0.005	0.036	0.013	0.104

[a] All results have been generated using completely uncontracted basis sets. The basis set limits have been taken as the aug-pcS-4 results. **KT3** and **B3LYP** denote the employed exchange-correlation functionals. ***H*** indicates hydrogen shielding constants (75 data points), ***M***$_1$ indicates Li and Be shielding constants (10 data points), ***A***$_1$ indicates C, N, O, and F shielding constants (92 data points), ***M***$_2$ indicates Na and Mg shielding constants (10 data points), and ***A***$_2$ indicates Si, P, S, and Cl shielding constants (32 data points).

The results in Table 5 have been grouped according to a qualitative classification corresponding to subdouble-, double-, triple-, quadruple-, and pentuple-ζ quality, which in most cases also indicates the highest angular momentum functions included in the basis sets (Table 2). From a computational point of view the total number of (contracted) functions is an important factor, and we have in Table 5 included the average number of basis functions per atom (N_{basis}) over the whole data set in Table 3 as an indicator of the computational cost. Since addition of both diffuse and tight functions rapidly increases the size of a basis set, the highest angular momentum functions included are not necessarily good indicators of the basis set size. The aug-cc-pCVTZ basis set, for example, contains more functions that the cc-pVQZ basis set, despite the latter formally being of higher ζ quality.

The unpolarized pcS-0 and aug-pcS-0 basis sets are not expected to be able to generate useful results, and basis sets of double-ζ quality are normally considered as the first level where semiquantitative results can be expected, with the 6-31G(d,p) basis set being widely used for routine applications. The performance of the 6-31G(d,p), 6-31++G(d,p), cc-pVDZ, and aug-pVDZ basis sets are very similar, with typical deviations of ~20 and ~30 ppm for first- and second-row elements, respectively. Addition of tight functions (aug-cc-pCVDZ) reduces the errors slightly but deteriorates the results for hydrogen. The Ahlrichs SVP basis set has somewhat larger errors. The aug-pcS-1 basis set clearly outperforms these standard basis sets and reduces the average error to ~3 ppm for first-row elements. The results for second-row elements are also improved to a value of ~15 ppm. The IGLO-II results formally compare favorably with the pcS-1 results, but it should be noted that the former results do not include the metallic systems, where many of the large deviations are observed. Without these systems, the MAD value of 8.2 ppm for the A_1 systems is reduced to 5.1 ppm, which can be compared with the IGLO-II value of 8.8 ppm.

At the triple-ζ level, the standard 6-311G(2df,2pd), 6-311++G(2df,2pd), cc-pVTZ, aug-cc-pVTZ, and Ahlrichs TZV basis sets perform roughly at par, with typical errors of ~0.1 ppm for hydrogen and ~10 ppm for the nonmetal atoms. The aug-pcS-2 basis set for comparison has errors of 0.03 ppm for hydrogen and less than 2 and 4 ppm for first- and second-row elements, respectively. The aug-cc-pCVTZ basis set improves the aug-cc-pVTZ results, showing that tight functions are important but also increases the size of the basis set by nine functions per atom on the average. Despite having significantly fewer functions, the aug-pcS-2 basis set performs better than the aug-cc-pCVTZ basis set. The IGLO-III basis set belongs to the triple-ζ family in terms of number of functions per atom but displays larger errors than the aug-pcS-2 basis set, despite the fact that the results do not include some of the more difficult systems.

The cc-pVQZ, aug-cc-pVQZ, and aug-cc-pCVQZ results are only marginally improved over those for the corresponding triple-ζ basis sets, and the Ahlrichs QZV basis set also only provides a small reduction in the basis set error compared to the TVZ basis set. In contrast, aug-pcS-3 reduces the basis set error by almost an order of magnitude relative to aug-pcS-2, and the mean error is now below 0.002 ppm for hydrogen and below 0.5 ppm for all the remaining elements.

At the pentuple-ζ level, the cc-pV5Z basis set only provides a minor improvement relative to the cc-pVQZ results, and there are still errors of ~15−20 ppm for the Na and Mg systems. The pcS-4 basis set, on the other hand, reduces the basis set errors to below 0.001 ppm for hydrogen and below ~0.1 ppm for all the remaining elements.

The grouping in Table 5 displays a significant variation in the number of basis functions per atom within each quality

Table 5. Mean Absolute Deviations (ppm) Relative to the Basis Set Limit for the Symmetry-Unique Nuclear Magnetic Shielding Constants for the Systems in Table *3*[a]

		KT3					B3LYP				
basis set	⟨N_{basis}⟩	H	M_1	A_1	M_2	A_2	H	M$_1$	A_1	M_2	A_2
pcS-0	6.4	1.5	10.8	63.9	6.6	92.9	1.5	10.4	63.6	9.9	100.0
aug-pcS-0	9.1	1.1	7.6	14.9	15.3	63.1	1.1	8.4	16.2	24.5	67.0
STO-3G	3.9	2.4	5.0	88.2	19.5	194.1	2.4	5.9	96.9	23.2	224.8
pcS-1	12.1	0.20	1.4	8.2	2.2	19.5	0.21	4.0	8.4	2.6	21.8
aug-pcS-1	18.6	0.08	1.4	3.3	1.3	15.9	0.09	3.2	3.9	2.9	21.0
cc-pVDZ	11.4	0.35	5.6	19.7	20.5	34.1	0.35	8.4	26.1	27.6	58.5
aug-cc-pVDZ	18.2	0.19	4.3	18.6	19.3	29.4	0.19	5.8	24.8	31.7	55.8
aug-cc-pCVDZ	20.7	0.22	3.4	12.8	14.4	18.3	0.20	3.8	17.7	20.7	25.4
6-31G(d,p)	11.3	0.47	5.0	23.8	16.9	30.4	0.43	7.4	29.5	24.1	49.7
6-31++G(d,p)	14.0	0.25	4.1	22.2	19.5	30.6	0.23	6.3	28.6	31.3	49.6
SVP	10.2	0.45	6.5	32.5	13.4	47.6	0.47	8.4	33.0	21.6	73.0
IGLO-II	16.5			(8.8)		(13.1)			(6.5)		(14.0)
pcS-2	25.0	0.07	0.7	3.5	1.0	6.3	0.07	2.2	2.3	2.1	6.8
aug-pcS-2	37.5	0.02	0.7	1.7	0.8	3.3	0.03	0.8	0.9	2.1	4.1
cc-pVTZ	24.4	0.19	2.2	10.5	11.2	20.5	0.19	2.5	9.5	17.9	26.1
aug-cc-pVTZ	37.3	0.10	2.1	10.6	15.7	18.6	0.11	2.1	9.5	13.5	22.1
aug-cc-pCVTZ	46.0	0.11	0.8	3.8	4.7	3.3	0.11	1.0	4.9	4.6	3.8
6-311G(2df,2pd)	24.4	0.20	2.2	10.3	0.8	11.4	0.20	2.2	8.8	1.1	14.6
6-311++G(2df,2pd)	27.1	0.10	2.2	9.4	0.9	5.4	0.11	2.2	7.8	1.6	8.2
TVZ	23.1	0.16	3.5	9.1	5.3	9.2	0.15	3.7	8.2	2.2	14.9
IGLO-III	24.9			(1.4)		(12.4)			(1.8)		(13.5)
pcS-3	54.7	0.004	0.05	0.37	0.22	0.97	0.005	0.05	0.22	0.43	1.02
aug-pcS-3	75.2	0.002	0.04	0.12	0.12	0.45	0.002	0.04	0.06	0.39	0.65
cc-pVQZ	45.5	0.083	0.74	7.31	13.30	26.90	0.082	0.42	4.23	15.43	35.99
aug-cc-pVQZ	66.6	0.051	0.73	6.71	18.12	25.41	0.052	0.41	3.81	5.92	33.41
aug-cc-pCVQZ	85.8	0.046	0.21	1.67	1.97	1.29	0.048	0.22	1.68	1.72	0.52
QVZ	45.9	0.065	1.08	5.92	3.08	3.33	0.066	1.28	3.39	0.79	3.20
pcS-4	92.3	0.001	0.01	0.08	0.03	0.15	0.001	0.02	0.04	0.21	0.13
cc-pV5Z	76.7	0.030	0.30	3.79	19.24	3.62	0.029	0.05	1.20	15.75	2.73

[a] All results using contracted basis sets. The basis set limit has been taken as the uncontracted aug-pcS-4 results. The cc-pVXZ and aug-cc-pVXZ basis sets include an additional tight d function for the elements Si−Cl. ⟨N_{basis}⟩ denotes the average number of basis functions per atom for the whole data set. KT3 and B3LYP denote the employed exchange-correlation functionals. *H* indicates hydrogen shielding constants (75 data points), M_1 indicates Li and Be shielding constants (10 data points), A_1 indicates C, N, O, and F shielding constants (92 data points), M_2 indicates Na and Mg shielding constants (10 data points), and A_2 indicates Si, P, S, and Cl shielding constants (32 data points). The IGLO basis sets are not defined for s-group elements, and the values in parentheses are for only 82 (A_1) and 26 (A_2) data points.

level: the aug-cc-pCVQZ basis set, for example, has 10 more basis functions than the aug-pcS-3 basis set, despite both being of quadruple-ζ quality augmented with tight and diffuse functions. In order to provide an alternative comparison, we have displayed the mean average deviation for all the non-hydrogen shielding constants as a function of the average number of functions per atom in Figure 1. The pcS-*n* and aug-pcS-*n* families of basis sets clearly display a smooth, controlled, and exponential convergence toward the limiting value. The cc-pVXZ and aug-cc-pVXZ basis sets display little convergence as the basis set is enlarged and have problems reducing the average error below 10 ppm. Augmenting with tight functions (aug-cc-pCVXZ) improves the results, but at a high computational cost, as many tight functions are added. It can be noted that the aug-cc-pCVTZ basis set has a much better performance than the cc-pVQZ set, despite the two basis sets being of almost the same size. The Ahlrichs- and Pople-type basis sets perform roughly at par with the cc-pVXZ basis sets, and augmenting the Pople basis sets with diffuse functions has only a small influence. Considering Figure 1, we find it significant that the aug-pcS-1 basis set, which is only marginally larger in size than the very popular 6-31++G(d,p), has basis set errors that are almost an order of magnitude smaller. Similarly, the aug-pcS-2 basis set, which is similar in size to the 6-311++G(2df,2pd) set, reduces the basis set errors by almost an order of magnitude.

When the MAD value is displayed as a function of the average number of functions per atom as in Figure 1, the convergence of the pcS-*n* and aug-pcS-*n* curves is seen to be very similar, with the aug-pcS-*n* results for a given *n* being of intermediate quality compared to the corresponding pcS-*n* and pcS-(*n*+1) results. When viewed in this fashion, the effect of augmenting the pcS-*n* basis sets with diffuse functions can be considered as simply being the results of having a more complete basis set. It can also be noted that the error reduction by including diffuse functions in Table 5 mainly arises from the polar systems in Table 3. For the nonpolar systems, which include a large fraction of typical organic molecules, the pcS-*n* basis sets provide results of quality similar to that for the aug-pcS-*n* basis set.

When statistical methods are used for evaluating the performance of various methods and basis sets, there is always a risk of biasing the results by the selection of

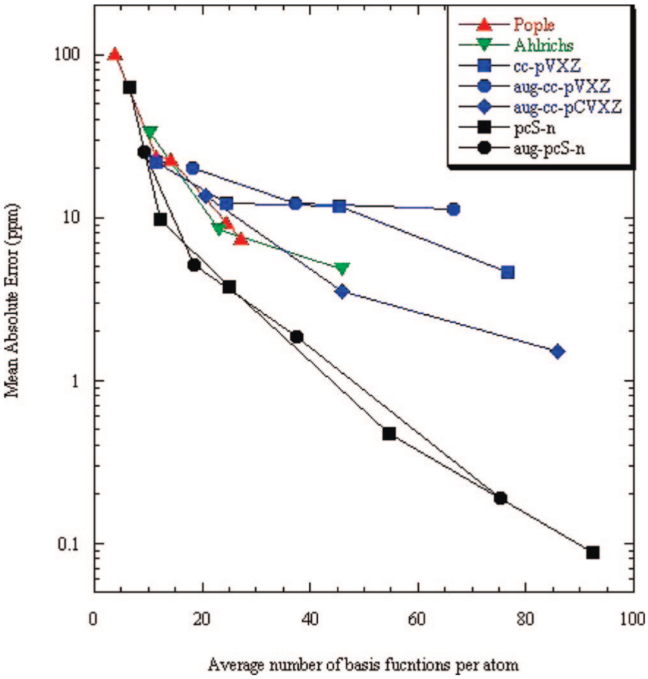


Figure 1. Mean absolute deviation relative to the basis set limit of nuclear magnetic shielding constants (ppm) for all non-hydrogen atoms as a function of the average number of basis functions per atom.

Table 6. Nuclear Magnetic Shielding Constants (ppm) for MgO

	KT3		B3LYP	
basis set	Mg	O	Mg	O
STO-3G	1686	−2727	4674	−13630
6-31G(d,p)	890	−1315	1046	−2966
6-31++G(d,p)	901	−1420	1018	−2610
6-311G(2df,2pd)	915	−1368	1046	−2766
6-311++G(2df,2pd)	918	−1359	1028	−2572
SVP	1990	−8587	−2275	23602
TVZ	915	−1356	1053	−2733
QVZ	908	−1344	1004	−2460
cc-pVDZ	880	−1310	1044	−3090
cc-pVTZ	887	−1372	1018	−2876
cc-pVQZ	892	−1402	996	−2712
cc-pV5Z	914	−1349	999	−2533
aug-cc-pVDZ	898	−1488	1029	−2889
aug-cc-pVTZ	898	−1385	992	−2586
aug-cc-pVQZ	888	−1385	980	−2591
aug-cc-pCVDZ	904	−1456	1026	−2773
aug-cc-pCVTZ	901	−1339	993	−2427
aug-cc-pCVQZ	903	−1335	996	−2424
pcS-0	1308	−3862	7336	−51269
pcS-1	1100	−2567	1484	−6031
pcS-2	921	−1436	1041	−2734
pcS-3	910	−1361	1008	−2477
pcS-4	906	−1338	1002	−2440
aug-pcS-0	916	−1476	1056	−2999
aug-pcS-1	937	−1525	1073	−2944
aug-pcS-2	922	−1439	1033	−2660
aug-pcS-3	909	−1354	1007	−2466
aug-pcS-4	907	−1343	1002	−2440

compounds in the test set. The systems in Table 3 were selected to represent a variety of structural elements found in many applications and covering a range of molecular bonding. In our initial selection of systems, we also had included the compounds BeO, BeS, MgO, and MgS. These systems, however, turned out to display pathological behaviors with respect to basis set convergence, and the errors were so large that they would have completely dominated the statistical measure in Tables 4 and 5 had they been included. The worst of these cases is the MgO system, for which the calculated shielding constants are shown in Table 6. The oxygen shielding constant with the B3LYP functional is calculated to be −2440 ppm with the (uncontracted) aug-pcS-4 basis set, and this value is presumably converged to within ~10 ppm, as judged from the aug-pcS-3 result. Using the −2440 ppm value as the reference, it is seen that all the basis sets of double-ζ quality have errors measured in the hundreds or thousands of ppm, and the SVP basis set marks a spectacular failure with a calculated value of +23 602 ppm. Part of this discrepancy is due to the fact that this basis set does not have d-type functions on Mg, but the pcS-1 result of −6031 ppm shows that this is not the only reason, as both these basis sets have the same angular momentum functions (Table 2). Basis sets of triple-ζ quality have typical errors of ~300 ppm, while augmentation with diffuse functions reduces the error to ~150 ppm. Only at the quadruple-ζ level, preferably augmented with both tight and diffuse functions, does the error drop to acceptable levels. For this specific system, inclusion of both d- and f-type functions on both atoms, as well as diffuse functions, is required to produce a qualitatively correct description. The aug-pcS-*n* basis sets here provide less accurate results than the aug-cc-pCVXZ basis set at the same ζ level. This is due to differences in how the higher angular momentum functions are included for Mg. The cc-pVXZ basis sets include d functions for Mg at the DZ level and f functions at the TZ level, analogous to the case for the p-block elements. The pcS-*n* basis sets, in contrast, only include p-type polarization functions at the DZ level (pcS-1) for s-block elements such as Mg and only up to d functions at the TZ level (pcS-2). It is therefore necessary to go to the pcS-3 level to obtain a qualitatively correct result.

The B3LYP oxygen shielding constant is the most sensitive to the quality of the basis set, but the same trend is seen for the magnesium atom. The KT3 functional provides similar trends, although the changes with respect to basis sets are less dramatic. The calculated shielding constants at the basis set limit differ by ~1000 ppm for oxygen and by ~100 ppm for magnesium between the two functionals, indicating the importance of selecting a suitable exchange-correlation functional. While double- or triple-ζ quality basis sets will be sufficient for the large majority of routine applications, it is in our opinion valuable to have a well-defined hierarchy of basis sets for systematically approaching the basis set limit for problematic cases, as for example MgO. It is gratifying to see that both the pcS-*n* and aug-pcS-*n* basis sets display a monotonic convergence toward the limiting value for this difficult system.

V. Summary

On the basis of our previous analysis for nuclear spin−spin coupling constants, we show that an improved basis set convergence for nuclear magnetic shielding constants can be obtained by addition of a single tight p-type basis function.

When used in combination with the previously proposed polarization consistent basis sets, this leads to the definition of a hierarchy of basis sets denoted pcS-*n*. An evaluation of the performance for a selection of typical molecular systems shows that these new basis sets represent an improvement with respect to reducing basis set errors relative to existing basis sets. A typical error at the aug-pcS-1 (double-ζ) level is 5 ppm for non-hydrogen atoms, which is reduced to 2 ppm upon going to the aug-pcS-2 (triple-ζ) level. The nuclear shielding constant for hydrogen displays a much smaller basis set effect, with typical errors of 0.1 and 0.03 ppm at the aug-pcS-1 and aug-pcS-2 levels.

Basis set limitations are only one possible error component in a comparison with experimental values, as the reference geometry, vibrational averaging, solvent effects,[37] and inadequacies in the exchange-correlation functional will need to be addressed in order to provide a direct comparison with experiments. The present pcS-*n* basis sets, however, should be suitable for controlling the basis set error, and the pcS-1 and pcS-2 basis sets should be suitable for Hartree−Fock and density functional methods in general and allow calculations for large systems.

The present work adds yet another sequence of basis sets to an already large variety, and it is reasonable to ask whether this represents an improvement of the computational capabilities or only serves to further complicate the selection of a basis set for a given problem. Not surprisingly, we favor the first option. Modern computational chemistry should in our opinion be able to control and assess the errors in the calculated results. An essential component for this is a well-defined hierarchy of basis sets which approaches the basis set limiting value in a smooth fashion and preferably is available for a reasonable selection of elements. Basis sets such as 6-31G(d,p) do not have a clear protocol for systematic improvements and must therefore be used as part of a precalibrated procedure, where the error evaluation is done by comparison with external reference data. Such an approach becomes problematic on encountering pathological cases, as illustrated by the MgO system in the present case, and for systems where no experimental data are available for calibration. The availability of a hierarchy of basis sets allows identification of pathological systems and provides the possibility of controlling the basis set errors, albeit at an increased computational cost.

The use of basis sets designed for specific properties has a long history,[38] but the pcJ-*n* and pcS-*n* basis sets in our opinion are the first to allow a systematic and fast convergence toward the basis set limiting value for nuclear magnetic properties. When property-specific basis sets are discussed, it should be recognized that basis sets are always a compromise between accuracy and computational efficiency. A basis set suitable for calculating a range of properties accurately will be so large that it is not computationally efficient. From an application point of view, the interest is usually in a single or narrow range of molecular properties, and having computationally efficient basis sets is necessary for tackling many real-world problems having a large number of atoms. The present pcS-*n* basis sets can be considered as a subset of the pcJ-*n* basis sets, where it is seen that nuclear magnetic shielding constants do not need as many tight functions as spin−spin coupling constants and can be contracted substantially harder without losing accuracy, thereby improving the computational efficiency. As such, we feel that the pcS-*n* basis sets should be a useful addition to the field of computational chemistry.

Acknowledgment. This work was supported by grants from the Danish Center for Scientific Computation and the Danish Natural Science Research Council.

Supporting Information Available: Tables giving exponents and contraction coefficients for the pcS-*n* and aug-pcS-*n* basis sets for the elements H−Ar. The information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lutnaes, O. B.; Ruden, T. A.; Helgaker, T. *Magn. Reson. Chem.* **2004**, *42*, S117. Helgaker, T.; Lutnaes, O. B.; Jaszunski, M. *J. Chem. Theory Comput.* **2007**, *3*, 86.

(2) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1995**, *103*, 3561. Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1996**, *104*, 2574. Stanton, J. F.; Gauss, J.; Siehl, H.-S. *Chem. Phys. Lett.* **1996**, *262*, 183.

(3) Parr, R. G.; Yang, W. *Density Functional Theory*; Oxford University Press: Oxford, U.K., 1989. Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, Germany, 2000.

(4) Ochsenfeld, C.; Kussmann, J.; Koziol, F. *Angew. Chem., Int. Ed.* **2004**, *43*, 4485. Schaller, T.; Buechele, U. P.; Klaerner, F.-G.; Blaeser, D.; Boese, R.; Brown, S. P.; Spiess, H. W.; Koziol, F.; Kussmann, J.; Ochsenfeld, C. *J. Am. Chem. Soc.* **2007**, *129*, 1293. Zienau, J.; Kussmann, J.; Koziol, F.; Ochsenfeld, C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4552.

(5) Keal, T. W.; Tozer, D. *J. Chem. Phys.* **2005**, *121*, 5654.

(6) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007. Wilson, A. K.; van Mourik, T., Jr *J. Mol. Struct.* **1996**, *388*, 339.

(7) Jensen, F. *J. Chem. Phys.* **2001**, *115*, 9113; **2002**, *116*, 3502; **2002**, *116*, 7372. Jensen, F.; Helgaker, T. *J. Chem. Phys.* **2004**, *121*, 3462.

(8) Kutzelnigg, W.; Morgan, J. D. *J. Chem. Phys.* **1992**, *96*, 4484; **1992**, *97*, 8821. (E).

(9) Klopper, W.; Kutzelnigg, W. *J. Mol. Struct.* **1986**, *135*, 339. Christensen, K. A.; Jensen, F. *Chem. Phys. Lett.* **2000**, *317*, 400. Jensen, F. *Theor. Chem. Acc.* **2000**, *104*, 484. Schwenke, D. W. *J. Chem. Phys.* **2005**, *122*, 014107.

(10) Jensen, F. *J. Chem. Phys.* **2003**, *118*, 2459.

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

(12) Jensen, F. *J. Chem. Phys.* **2002**, *117*, 9234.

(13) Jensen, F. *J. Chem. Theory Comput.* **2006**, *2*, 1360.

(14) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293. Karadakov, P. B. *Mod. Magn. Reson.* **2006**, *1*, 59.

(15) Vaara, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399.

(16) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497. Zhang, Y.; Wu, A.; Xu, X.; Yan, Y. *J. Phys. Chem. A* **2007**, *111*, 9431. Wu, A.; Zhang, Y.; Xu, X.; Yan, Y. *J. Comput. Chem.* **2007**, *28*, 2431. d'Antuono, P.; Botek, E.; Champagne, B.; Spassova, M.;

Denkova, P. *J. Chem. Phys.* **2006**, *125*, 144309. Chesnut, D. B. *Chem. Phys. Lett.* **2003**, *380*, 251.

(17) Wu, A.; Cremer, D.; Gauss, J. *J. Phys. Chem. A* **2003**, *107*, 8737. Magyarfalvi, G.; Pulay, P. *J. Chem. Phys.* **2003**, *119*, 1350. Auer, A. A.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2003**, *118*, 10407.

(18) Sefzik, T. H.; Turco, D.; Iuliucci, R. J.; Facelli, J. C. *J. Phys. Chem. A* **2005**, *109*, 1180. Jaszunski, M. *Chem. Phys. Lett.* **2004**, *385*, 122. Moon, S.; Case, D. A. *J. Comput. Chem.* **2006**, *27*, 825. Kupka, T.; Ruscic, B.; Botto, R. E. *J. Phys. Chem. A* **2002**, *106*, 10396. Kupka, T.; Ruscic, B.; Botto, R. E. *Solid State Nucl. Magn. Reson.* **2003**, *23*, 145.

(19) Kupka, T.; Lim, C. *J. Phys. Chem. A* **2007**, *111*, 1927.

(20) Schindler, M.; Kutzelnigg, W. *J. Chem. Phys.* **1982**, *76*, 1919. Kutzelnigg, W.; Fleischer, U.; Schindler, M. *The IGLO-Method: Ab Initio Calculation Interpretation of NMR Chemical Shifts Magnetic Susceptibilities*; Springer-Verlag: Heidelberg, Germany, 1990; 23.

(21) Manninen, P.; Vaara, J. *J. Comput. Chem.* **2006**, *27*, 434.

(22) Helgaker, T.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J.; Ruud, K.; Ågren, H.; Auer, A. A.; Bak, K. L.; Bakken, V.; Christiansen, O.; Coriani, S.; Dahle, P.; Dalskov, E. K.; Enevoldsen, T.; Fernez, B.; Hättig, C.; Hald, K.; Halkier, A.; Heiberg, H.; Hettema, H.; Jonsson, D.; Kirpekar, S.; Kobayashi, R.; Koch, H.; Mikkelsen, K. V.; Norman, P.; Packer, M. J.; Pedersen, T. B.; Ruden, T. A.; Sanchez, A.; Saue, T.; Sauer, S. P. A.; Schimmelpfenning, B.; Sylvester-Hvid, K. O.; Taylor, P. R.; Vahtras O., DALTON, a Molecular Electronic Structure Program, Release 2.0, 2005.

(23) 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; Gaussian, Inc., Wallingford, CT, 2004.

(24) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(25) London, F. *J. Phys. Radium, Paris* **1937**, *8*, 397. McWeeny, R. *Phys. Rev.* **1962**, *126*, 1028. Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789. Wolinski, K.; Hilton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.

(26) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 764. Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(27) Hehre, W. J.; Ditchfield, R.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1970**, *52*, 2769.

(28) Francl, M. M.; Petro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654.

(29) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265. Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *Theor. Chem. Acc.* **1980**, *72*, 650.

(30) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639.

(31) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358. Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.

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

(33) Woon, D. E. *J. Chem. Phys.* **1995**, *103*, 4572.

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

(35) Schafer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(36) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753.

(37) Cybulski, H.; Sadley, J. *Chem. Phys.* **2006**, *323*, 218. Aidas, K.; Moegelhoej, A.; Kjaer, H.; Nielsen, C. B.; Mikkelsen, K. V.; Ruud, K.; Christiansen, O.; Kongsted, J. *J. Phys. Chem. A* **2007**, *111*, 4199.

(38) Sadlej, A. J. *Chem. Phys. Lett.* **1977**, *47*, 50. Roos, B. O.; Sadlej, A. *J. Chem. Phys.* **1985**, *94*, 43. Wolinski, K.; Roos, B. O.; Sadlej, A. J. *Theor. Chim. Acta* **1985**, *68*, 431. Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995. Benkova, Z.; Sadlej, A. J.; Oakes, R. E.; Bell, S. E. J. *J. Comput. Chem.* **2005**, *26*, 145. Baranowska, A.; Siedlecka, M.; Sadlej, A. J. *Theor. Chem. Acc.* **2007**, *118*, 959.

CT800013Z

JCTC Journal of Chemical Theory and Computation

Ab Initio Highly Correlated Conformational Analysis of 1,2-Difluorethane and 1,2-Dichloroethane

Mauro L. Franco,[†,‡] Dalva E. C. Ferreira,[†] Hélio F. Dos Santos,[§] and Wagner B. De Almeida*,[†]

Laboratório de Química Computacional e Modelagem Molecular (LQC-MM), Departamento de Química, ICEx, Universidade Federal de Minas Gerais (UFMG), Campus Universitário, Pampulha, Belo Horizonte, MG, 31270-901, Brazil, Diretoria de Ciências Exatas (DCX), Centro Universitário do Leste de Minas Gerais, UniLeste-MG, Campus I, Coronel Fabriciano, MG, 35170-056, Brazil, and Núcleo de Estudos em Química Computacional (NEQC), Departamento de Química, ICE, Universidade Federal de Juiz de Fora (UFJF), Campus Universitário, Martelos, Juiz de Fora, MG, 36036-330, Brazil

Received January 14, 2008

Abstract: Temperature-dependent conformational population calculations for anti and gauche forms of 1,2-dichloroethane and 1,2-difluorethane were carried out at a highly correlated level of theory (MP4(SDTQ) and CCSD(T)) employing good quality basis sets (6-311++G(3df,3pd) and aug-cc-pVQZ) for the determination of gas relative conformational energies, making use of the statistical thermodynamics formalism for the evaluation of the thermal energy correction at the MP2/6-311++G(3df,3pd) and MP2/aug-cc-pVTZ levels. In addition to the standard calculation of thermodynamic partition functions, a treatment of the lowest-frequency vibrational mode as hindered rotation and anharmonic correction to vibrational frequencies was also included. We found a good agreement between ab initio calculated conformational population values and experimental gas-phase electron diffraction data for the 1,2-dicloroethane. However, for the 1,2-difluorethane species, a reasonable agreement with the experimental anti/gauche population ratio obtained from the analysis of gas-phase far-infrared (50–370 cm^{-1}) and low-frequency Raman (70–300 cm^{-1}) spectra was not obtained. The results reported here indicate that, for 1,2-difluorethane, and probably other substituted alkanes where the gauche effect is of relevance, a more appropriated treatment of the low-frequency modes must be pursued in order to reproduce experimental conformational population data.

1. Introduction

Conformational analysis is a fascinating subject, first related to organic chemistry, which has attracted the attention of experimentalists and theoreticians for a long time, being also of great importance to almost all areas of chemistry. Over the past years, temperature-dependent nuclear magnetic resonance (NMR) and solvent effect studies have been reported by many researchers in the area of organic chemistry; a recent example is the conformational analysis of succinic acid.[1] In several cases, the conformational process is not simple, with some vibrational modes being associated with small rotational barriers around C—C single bonds. Therefore, the rationalization of the governing factors operating on nonrigid molecules is still not completely clear, even for small systems, such as alkane-substituted molecules.[2] The substituted ethane molecules, such as 1,2-dicloroethane[3–5] and 1,2-difluorethane,[3,6–8] have been the subject of a considerable number of investigations motivated by the interest in its restricted internal rotation.

* Corresponding author fax: +55 31 34995700, e-mail: wagner@netuno.qui.ufmg.br.

[†] UFMG.

[‡] Centro Universitário do Leste de Minas Gerais.

[§] UFJF.

10.1021/ct800014w CCC: $40.75 © 2008 American Chemical Society
Published on Web 03/22/2008

In addition, the recent literature for the simple nonsubstituted ethane molecule also shows that attention has been paid to the reason for the rotational barrier leading to the experimentally observed staggered structure.[9,10] It is well-known that for 1,2-dichloroethane the anti form predominates over the gauche conformer. However, the opposite is observed for 1,2-difluoroethane, where both experimental and theoretical investigations have shown that this molecule prefers the gauche conformation, which has been successfully rationalized in terms of a hyperconjugation model.[11] So, in the case of the 1,2-difluoroethane molecule, the stability of the gauche conformation has been attributed to the high electronegative character of the fluorine atom denominated the *gauche effect*, where the equilibrium geometry is a result of charge transfer from the C−H electron to the C−F* antibonds.[12] Investigation of the far-infrared (50–370 cm^{-1}) and low-frequency Raman (70–300 cm^{-1}) spectra[8] of the gas-phase sample of 1,2-difluoroethane showed that the gauche conformer is 3.39 ± 0.54 kJ mol^{-1} more stable than the anti form, and it has been one of the most discussed cases of intramolecular interaction over the past three decades.

Analogously, the gauche preference of 1-fluoropropane[12] has also been recognized as of hyperconjugative nature. On the other side for butane,[13,14] the anti preference has been attributed to steric hindrance or solvent effects. Freitas and Rittner[2] evaluated the conformational behavior of 1,2-di-substituted ethanes, where the bulky substituents are used, such as CN and NO_2 groups. They have shown that the electron delocalization contributes very differently to the conformational equilibrium of the 1,2-di-substituted ethanes. For example, for 1,2-dinitroethane, the contribution from electronic delocalization strongly favored the anti conformation, though antiperiplanar C−H/C−N(O2) hyperconjugative, analogous to the interaction usually taken as the driving force of the gauche effect in fluorinated compounds, also showed a high energy. Although studies with model compounds search to describe the origin of the gauche effect, there is no general rule for this effect in the conformational isomerism of 1,2-di-substituted ethanes.[2,15]

Concomitantly with the interest in the anti/gauche population ratio of substituted ethane, the ethane molecule[9,16,17,4,10,11] has also been the subject of a considerable number of investigations addressing restricted internal rotation. The gas-phase spectroscopic and thermodynamic experimental data reported for ethane and ethane-substituted molecules provide useful information to assess the capability of available theoretical methods used to calculate temperature-dependent macroscopic properties. The theoretical determination of thermodynamic properties and, so, gas-phase conformational population (Gibbs population ratio) is based on the use of quantum mechanical methods and the standard statistical thermodynamics formalism,[18] where thermodynamic quantities are calculated through the evaluation of electronic, translational, rotational, and vibrational molecular partition functions. The presence of low-frequency modes, which are not true vibrations, presents a challenge for the evaluation of the vibrational partition function that so far has no general solution, and therefore approximated treatments have been proposed. Various theoretical models have been developed to account for the internal rotation problem. The most used recently reported treatments can be found, for example, in refs 19 and 20.

In order to investigate the performance of theoretical approaches for predicting relative gas-phase conformational population values, as compared to observed experimental data, two distinct points must be considered: the adequacy of the theoretical model employed, which is reflected in the pertinence of the mathematical equations developed, and the quality of the calculated energy values used to feed the mathematical functions to produce numerical values for the population ratio, which is dictated by the ab initio level of theory employed. Regarding the calculation of Gibbs conformational population we have, on one hand, the statistical thermodynamic formalism which makes use of molecular partition functions based on Boltzmann distributions and also additional corrections for hindered rotation through the use of empirical formulas, and on the other hand, the quantum mechanical methods available for the resolution of the time-independent Schrödinger equation for an isolated molecule in the vacuum, which produce the various energy values (electronic, rotational, and vibrational) and structural data to feed the thermodynamic partition functions. At this point, the validity of the theoretical approaches is attested to by comparison with experimental conformational population data within experimental uncertainties.

In the light of the importance of the gauche effect in conformational analysis studies, as mentioned in paragraphs before, we decided to investigate two experimentally well-known cases where, in one of them, this effect is of relevance: 1,2-dichloro- and 1,2-difluorethane. Then, through comparison with available gas-phase experimental conformational population data, we aim to analyze the performance of the theoretical approaches commonly used to calculate conformational population values. In this work, we used the statistical thermodynamics formulas for the calculation of thermal energies employing ab initio post-Hartree−Fock (HF) quantum chemical methods to the calculation of geometrical parameters and harmonic frequencies required for the evaluation of rotational and vibrational partition functions.[18] The theoretical model used in this article to describe the thermal energy corrections also included a hindered rotation treatment[19] of low-frequency modes and anharmonicity correction[21,22] to the vibrational frequencies for the evaluation of the vibratonal partition function.

2. Computational Methodology

The equilibrium geometry of the two minimum-energy conformers (anti and gauche) found on the electronic ground-state potential energy surface (PES) for the 1,2-dicloroethane and 1,2-difluorethane molecules were fully optimized with no symmetry constraint or other geometrical restriction at the second-order Møller−Plesset perturbation theory (MP2)[23] using Pople's split-valence 6-311++G(3df,3pd) basis set[24] and Dunning's correlated consistent basis set (aug-cc-pVTZ).[25] In addition, two first-order transition state (TS) structures connecting the anti and gauche conformers (named TS1) and two equivalent gauche forms (named TS2) were also located on the PES; therefore, energy barriers for conformational interconversion could be calculated. The equilibrium structures (and TS) were fully optimized and vibrational frequencies, and therefore thermodynamic quantities, calculated at the MP2/6-311++G(3df,3pd) and MP2/aug-cc-pVTZ levels of theory. In order to obtain a better description of the electronic correlation effects, single-point energy calculations, using MP2 fully optimized geometries, were also carried out at the fourth-order Møller−Plesset perturbation theory with single, double, triple, and quadruple excitations (MP4(SDTQ))[26] and coupled cluster with single, double, and perturbative triple excitations (CCSD(T)).[27] For 1,2-difluorethane, the G3(MP2) method[28] was also employed for the calculation of relative anti/gauche energies, in order to evaluate the gauche effect in the light of the G3(MP2) methodology as compared to the MP4(SDTQ) and CCSD(T) calculations.

The gas-phase thermal correction (ΔG_{T}) to the relative energy values ($\Delta E_{\text{ele-nuc}}$) was evaluated with the MP2/6-311++G(3df,3pd) and MP2/aug-cc-pVTZ structural parameters and harmonic vibrational frequencies and, then, used to obtain the gas-phase Gibbs free energy (ΔG) according to eq 1, where the double slash means that a MP4(SDTQ) or CCSD(T) single-point energy calculation was performed using the MP2 fully optimized geometry.

$$\Delta G = \Delta E_{\text{ele-nuc}}^{\text{MP4(SDTQ)//MP2,CCSD(T)//MP2}} + \Delta G_{\mathrm{T}}^{\mathrm{MP2}} \qquad (1)$$

$\Delta E_{\text{ele-nuc}}$ in eq 1 represents the electronic plus nuclear repulsion energy contribution, originated from the resolution of the time-independent Schrödinger equation for an isolated molecule in the perfect vacuum, and the second term is the temperature-dependent thermal energy correction, which is given by eq 2:

$$\Delta G_{\mathrm{T}} = \Delta U - T\Delta S \qquad (2)$$

where ΔU is the internal energy correction to enthalpy which includes the zero-point energy (ZPE) contribution and $T\Delta S$ is the entropic contribution at the absolute temperature T. The mathematical expression for the terms on the right side of eq 2 and the respective definition of the electronic, translational, rotational, and vibrational partition functions can be found in ref 18. In analogy to eq 1, we can write the relative enthalpy as

$$\Delta H = \Delta E_{\text{ele-nuc}} + \Delta U \qquad (3)$$

which gives directly $\Delta G = \Delta H - T\Delta S$, well-known from classical thermodynamics. The temperature range used was chosen according to the reported experimental data available (with $p = 1$ atm).

For the evaluation of the vibrational and rotational contributions to the thermal corrections given by eq 2, the harmonic oscillator (HO) and rigid rotor (RR) partition functions were used.[18] The anharmonic correction was evaluated according to the procedure discussed in refs 21 and 22, using a second-order perturbative treatment based on quadratic, cubic, and semidiagonal quartic force constants, which consists of retaining the formal expression of the HO partition function, but the ZPE and vibrational frequencies (ν_i) are obtained at the anharmonic level. Therefore, in this approach, only the vibrational frequencies are corrected for anharmonicity (the HO vibrational partition function was used). Explicit equations for anharmonic terms can be found in refs 21 and 22.

It has been well-known for well over half a century that a treatment of low-frequency vibrational modes, which are not true vibrations, as hindered rotations is required to describe the thermodynamics of ethane and ethane-substituted molecules. In ref 19, a treatment of low-frequency modes as internal hindered rotation is described in detail, with a procedure for the automatic identification of low-frequency modes as a hindered rotor, requiring no user intervention (implemented in the Gaussian computer code), being reported. Following early works of Pitzer and Gwinn[29] tabulating thermodynamic functions, formulas became available to interpolate the partition function between those of a free rotor, hindered rotor, and harmonic oscillators,[29–32] with the approximation by Truhlar[31] being used in many studies in recent years. In ref 19, a modified approximation to the hindered rotor partition function for the ith low-frequency mode (named here $q_i^{\text{Hind-Rot}}$) was given, which was used in this work. As stated in ref 19, this improved partition function keeps the good characteristics of the previous equation proposed by Pitzer and Gwinn,[29] while enhancing its behavior for low values of V_0/kT (V_0 is the barrier height for internal rotation, k the Boltzmann constant, and T the absolute temperature). These formulas (see ref 19) are for one normal vibrational mode involving a single rotating group with a clearly defined moment of inertia. The thermal corrections to enthalpy and Gibbs free energy, including hindered rotation and anharmonic correction to vibrational frequencies, are calculated according to eqs 4 and 5 below, at the MP2 level of theory and with good quality basis sets (6-311++G(3df,3pd) and aug-cc-pVTZ). The terms Hind-Rot and Anh indicate the use of hindered rotation and anharmonicity correction to vibrational frequency treatments, respectively, to account for deviations from the rigid rotor−harmonic oscillator (RR-HO) partition function.

$$\Delta U^{\text{Hind-Rot-Anh}} = \Delta U + \Delta U^{\text{Hind-Rot}} + \Delta U^{\text{Anh}} \qquad (4)$$

$$\Delta G_T^{\text{Hind-Rot-Anh}} = \Delta G_T + \Delta G_T^{\text{Hind-Rot}} + \Delta G_T^{\text{Anh}} \qquad (5)$$

All quantum chemical calculations were done with the Gaussian package,[33] where the hindered rotation treatment and anharmonicity correction calculations are readily implemented, at the Laboratório de Química Computacional e Modelagem Molecular, Departamento de Química, Universidade Federal de Minas Gerais, and also Núcleo de Estudos em Química Computacional, Departamento de Química, Universidade Federal de Juiz de Fora.

3. Results and Discussions

Table 1 reports the calculation of absolute entropy for ethane at room temperature, using the MP2 level of theory and a series of Pople's split-valence basis sets, with the aid of the standard statistical thermodynamics formalism with the inclusion of a treatment of the hindered-rotation effects and anharmonicity correction to vibrational frequencies, as explained in the methodology section. These data are shown only for reasons of comparison, since a number of theoretical

Table 1. MP2 Calculation of Absolute Entropy (J mol^{-1} K^{-1}) of the Ethane Molecule in the Staggered Form (T = 298 K, p = 1 atm) Employing Diverse Basis Sets

Entropy	6-31G (d,p)	6-311++G (d,p)	6-311++G (2d,2p)	6-311++G (2df,2pd)	6-311++G (3df,3pd)	aug-cc-pVTZ
S^{a}	226.65	227.11	227.27	227.23	227.15	227.32
$S^{\text{True-Vib}}$ [b]	221.50	221.79	221.67	221.75	221.71	221.75
$S^{\text{Hind-Rot-Anh}}$ [c]	228.07	228.74	228.91	228.86	228.86 {0.3%}[d]	229.03 {0.2%}[d]

[a] Absolute entropy calculated using the standard statistical thermodynamics partition function (particle in a box; rigid rotor and harmonic oscillator approximations for translational, rotational, and vibrational contributions) including all 3N-6 vibrational modes as harmonic oscillators. $S = S_{\text{trans}} + S_{\text{rot}} + S_{\text{vib}}$ (S_{trans} = 151.17 and S_{rot} = 68.03 J mol^{-1} K^{-1} (MP2/6-311++G(3df,3pd) value). 1 cal = 4.184 J. [b] The low-frequency mode was excluded from the evaluation of the vibrational partition function for the calculation of the absolute entropy, so 3N-7 normal modes were used. Only the true vibrational modes that can be satisfactorily described as harmonic oscillators were considered. The low-frequency contribution to entropy ($S_{\text{vib}}{}^{\text{Low-Freq}}$) is 5.44 J mol^{-1} K^{-1}. $S_{\text{vib}}{}^{\text{True-Vib}}$ = 2.51 J mol^{-1} K^{-1}; $S_{\text{vib}}{}^{\text{Low-Freq}}$ = 5.44 J mol^{-1} K^{-1}; $S_{\text{vib}}{}^{\text{Hind-Rot}}$ = 1.05 J mol^{-1} K^{-1}; $S_{\text{vib}}{}^{\text{Anh}}$ = 0.67 J mol^{-1} K^{-1}. $S_{\text{vib}}{}^{\text{Hind-Rot-Anh}} = S_{\text{vib}}{}^{\text{True-Vib}} + S_{\text{vib}}{}^{\text{Low-Freq}} + S_{\text{vib}}{}^{\text{Hind-Rot}} + S_{\text{vib}}{}^{\text{Anh}}$ = 9.67 J mol^{-1} K^{-1}. [c] Absolute entropy value calculated with the inclusion of anharmonicity and hindered internal rotation corrections for the evaluation of the vibrational partition function. [d] Percent error relative to the experimental entropy value (229.49 ± 0.8 J mol^{-1} K^{-1}) obtained at 298.1 K from ref 35. The corresponding errors for the TS value are only 0.17 and 0.13 kJ mol^{-1}, respectively, for the 6-311++G(3df,3pd) and aug-cc-pVTZ basis sets.

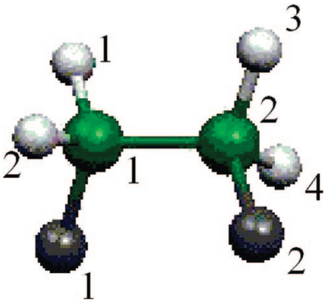
studies have been reported for the simple ethane molecule. When the values given in the last column of Table 1 are analyzed, it can be seen that the entropy calculated with the aug-cc-pVTZ basis set deviates only 0.17 J mol^{-1} K^{-1} from the value calculated with Pople's split valence basis set, with a smaller deviation obtained when the low-frequency mode is ignored (0.04 J mol^{-1} K^{-1}). From the results reported in Table 1, it can be seen that the combination of anharmonic correction to vibrational frequencies and a hindered rotor treatment of the lowest-frequency modes provides a perfect description of the entropy of ethane at room temperature, when a large basis set is used (at least 6-311++G(2d,2p)) with a MP2 calculation. The deviation from the experimental value is only 0.3% (0.2% with the aug-cc-pVTZ basis set), which is within the experimental uncertainty of ±0.63 J mol^{-1} K^{-1}. Therefore, for the ethane molecule, the approach used worked very well, an expected and well-known result in the literature. It is worth mentioning that the MP2/6-311++G(3df,3pd) total entropy (the sum of electronic, translational, rotational, and vibrational contributions is 228.86 J mol^{-1} K^{-1}) has a relatively small vibrational contribution of 9.67 J mol^{-1} K^{-1}, which, however, is assumed to play the key role when comparison with experimental data is made. The potential barrier for ethane ($V = {}^{1}/_{2}V_0(1 + \cos 3\varphi)$) was also calculated at the MP2, MP4(SDQ), MP4(SDTQ), CCSD, and CCSD(T) levels of theory (MP2-optimized geometries), with the 6-311++G(3df,3pd), aug-cc-pVTZ, and aug-cc-pVQZ basis sets (*staggered* → *eclipsed* process). We found a smooth dependence with the level of electron correlation for all methods showing a very reasonable agreement with experimental results within experimental uncertainties (the CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) value is 11.80 kJ mol^{-1} and the experimental value 12.03 ± 0.52 kJ mol^{-1}).

An assessment of the suitability of the MP2/6-311++G(3df,3pd) and MP2/aug-cc-pVTZ levels of calculation to evaluate geometrical parameters, rotational constants, and vibrational frequencies, required by the numerical calculation of partition functions, can be made by analyzing the data reported in Tables 2 and 3 for 1,2-difluorethane (as an example). The good agreement with experimental results shown in Table 2 for the rotational constants guarantees that the rotational partition function is satisfactorily well-represented at the MP2 level of theory. The same holds for the comparison between experimental and theoretical vibrational frequencies reported in Table 3. When anharmonicity correction is included, the deviation between experimental gas-phase vibrational frequencies and calculated anharmonic frequencies is below 50 cm^{-1} for the CH stretch vibrations (high frequencies), being even smaller for deformation and twist modes (less than 20 cm^{-1}) and lower frequency modes below 1000 cm^{-1} (less than 5 cm^{-1}).

Figure 1 shows calculated MP2 thermal quantities (ΔU and ΔG_{T}) using various basis sets for the anti → gauche process for 1,2-difluorethane (a similar behavior was found for 1,2-dichloroethane). It can be seen from Figure 1 that the thermal corrections reached nearly unchanged values within ±0.08 kJ mol^{-1} at the MP2/6-311++G(3df,3pd) level of theory for geometry optimization and harmonic frequency calculation (the same conclusion is reached with the aug-cc-pVTZ basis set), a variation that would cause a change on the calculated conformational population of less than 1%. It is worth saying that an MP4 or CC geometry optimization and frequency calculation with a triple-ζ-quality basis set for 1,2-difluorethane and 1,2-dicloroethane is unthinkable; even so, we would not expect a significant modification of the pattern exhibited in Figure 1. Just recently, we have performed MP4(SDQ) and CCSD geometry optimizations and harmonic frequency calculations, employing the 6-311++G(2d,2p) and aug-cc-pVDZ basis sets, for the CF_2Cl_2 species,[37] with the results being in excellent agreement with MP2 calculations, which adds confidence to our expectation regarding the trend shown in Figure 1.

The effect of the electronic correlation and size of the basis set on relative electronic plus nuclear repulsion energies ($\Delta E_{\text{ele-nuc}}$) can be analyzed from the results reported in Figure 2 for 1,2-difluorethane (a similar pattern was obtained for 1,2-dicloroethane). It can be seen that the MP4(SDTQ) and CCSD(T) relative energies for the anti → gauche process agree within less than 0.21 kJ mol^{-1}, showing a welcome smooth behavior of the energy values as a function of the level of theory and basis set quality. It can also be seen from Figure 2 that the difference between the CCSD(T)/6-311++G(3df,3pd) and CCSD(T)/aug-cc-pVQZ relative energies is less than 0.08 kJ mol^{-1}, so the use of MP2/6-311++G(3df,3pd) thermal corrections and CCSD(T)/6-311++G(3df,3pd) relative energies can be justified. We may

Table 2. *Ab Initio* and Experimental Geometrical Parameters (the Atomic Labels Are Defined Below) and Rotational Constants for the 1,2-Difluorethane gauche Conformer[a]



	exptl.	6-31G(d)	6-311++G(2d,2p)	6-311++G(3df,3pd)	aug-cc-pVDZ	aug-cc-pVTZ
			Bond Distances (Å)[b]			
C_1-C_2	1.493 ± 0.002	1.501	1.498	1.499	1.507	1.499
C_1-F_1 ($=C_2-F_2$)	1.390 ± 0.003	1.392	1.392	1.383	1.407	1.388
C_1-H_1 ($=C_2-H_3$)	1.099 ± 0.002	1.095	1.088	1.090	1.102	1.091
C_1-H_2 ($=C_2-H_4$)	1.093 ± 0.004	1.093	1.086	1.088	1.100	1.089
			Bond Angles (deg)[b]			
$F_1-C_1-C_2$ ($=F_2-C_2-C_1$)	110.6 ± 0.5	109.5	110.0	110.3	110.3	110.3
$H_1-C_1-C_2$ ($=H_3-C_2-C_1$)	108.4 ± 0.6	110.3	110.7	110.6	109.7	110.7
$H_2-C_1-C_2$ ($=H_4-C_2-C_1$)	113.3 ± 0.6	110.3	109.9	109.7	111.0	109.7
$F_1-C_1-H_1$ ($=F_2-C_2-H_3$)	109.6 ± 0.3	108.6	108.0	108.1	107.8	108.0
$F_1-C_1-H_2$ ($=F_2-C_2-H_4$)	107.8 ± 0.6	108.5	108.1	108.2	107.7	108.2
$H_1-C_1-H_2$ ($=H_3-C_2-H_4$)	109.1 ± 0.5	109.6	110.1	109.9	110.3	110.0
			Torsion Angle (deg)[b]			
F−C−C−F	71.0 ± 0.3	68.9	70.5	70.3	70.7	70.8
			Rotational Constants (MHz)[b]			
A	17322	16855[c]	17158	17349	16871	17295
B	5013	51448[c]	5062	5073	4954	5040
C	4383	44505[c]	4409	4424	4318	4400

[a] Theoretical values were obtained using the MP2 level and employing various basis sets. [b] Microwave values taken from ref 36. Symmetry number: $\sigma = 2$. [c] The corresponding MP2/6-311G(d,p) *A*, *B*, and *C* values are respectively 17234, 5060, and 4410 MHz.

Table 3. Experimental and *ab Initio* MP2 Vibrational Frequencies for the 1,2-Difluorethane Gauche Conformer

	exptl.	MP2/6-311++G (2d,2p)		MP2/6-311++G (3df,3pd)	MP2/aug-cc-pVDZ	MP2/aug-cc-pVTZ
vibrational frequencies (cm^{-1})[a]	observed frequencies	anharmonic frequencies[b]	harmonic oscillator	harmonic oscillator	harmonic oscillator	harmonic oscillator
FCCF torsion	147	151	153	154	151	154
CCF bend	327	326	326	328	319	326
CCF bend	500	498	501	506	491	502
CC stretch	865	868	885	892	874	886
CH_2 rock	896	895	914	920	893	915
CF stretch	1076	1066	1090	1109	1071	1101
CF stretch	1079	1095	1122	1142	1110	1135
CH_2 rock	1116	1123	1148	1150	1139	1146
CH_2 twist	1244	1251	1278	1278	1250	1274
CH_2 twist	1284	1296	1322	1323	1296	1318
CH_2 wag	1377	1393	1427	1418	1388	1415
CH_2 wag	1410	1425	1462	1457	1435	1453
CH_2 deformation	1460	1476	1516	1513	1480	1513
CH_2 deformation	1460	1478	1517	1513	1484	1513
CH_2 symmetric stretch	2958	2997	3105	3094	3094	3091
CH_2 symmetric stretch	2985	3004	3114	3101	3102	3098
CH_2 antisymmetric stretch	2995	3030	3173	3163	3169	3158
CH_2 antisymmetric stretch	3001	3042	3185	3174	3179	3169

[a] Experimental values and assignments obtained from a gas-phase infrared and Raman study reported in ref 8. [b] Evaluated including anharmonic corrections.

say that the MP4(SDTQ) and CCSD(T) conformational energies might be trusted with a rough uncertainty estimated at ±0.21 kJ mol^{-1}, based on the pattern shown in Figure 2, with a corresponding uncertainty in the conformational population of approximately 1%. Nevertheless, as will be shown later, this 1% of uncertainty cannot be blamed when a comparison with experimental results is made. The reported uncertainties for experimental conformational populations are in the range of ±2–5%, and the uncertainty value for experimental enthalpy determination is within ±0.4–0.8 kJ

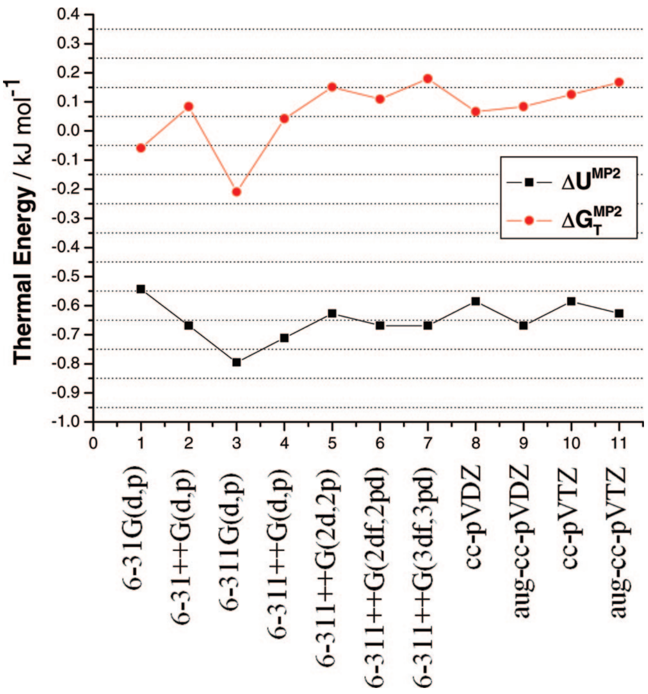


Figure 1. Anti → gauche MP2 thermal energy (at room temperature) variation for 1,2-difluorethane as a function of the basis set quality. The MP2/6-311++G(3df,3pd) and MP2/aug-cc-pVTZ $T\Delta S$ (entropic contribution) values are respectively –0.84 and –0.80 kJ mol^{-1} ($\Delta G_T = \Delta U - T\Delta S$).

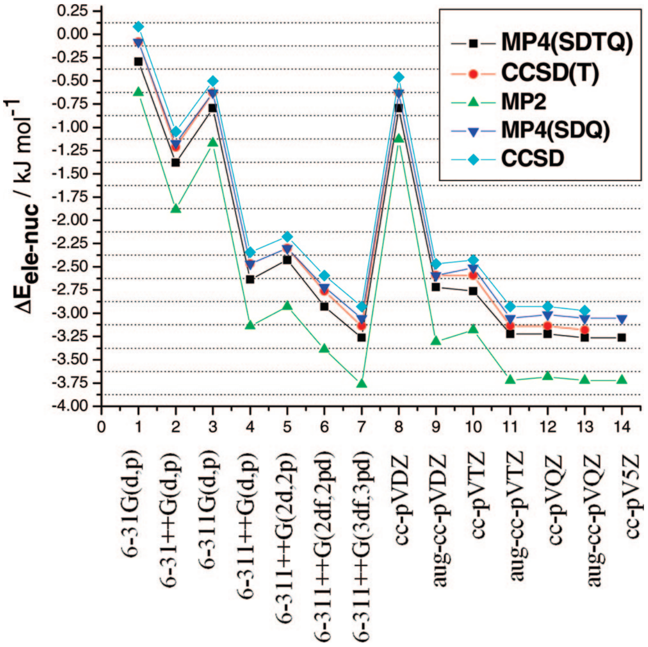


Figure 2. Anti → gauche energy ($\Delta E_{\text{ele-nuc}}$ in the vacuum) variation for 1,2-difluorethane as a function of the level of calculation. The CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) and CCSD(T)/aug-cc-pVQZ//MP2/aug-cc-pVTZ relative energy values are respectively –3.14 and –3.18 kJ mol^{-1}. The corresponding MP4(SDTQ) values are respectively –3.26 and –3.26 kJ mol^{-1} (the MP4(SDTQ)/cc-pV5Z//MP2/aug-cc-pVTZ value is –3.26 kJ mol^{-1}).

mol^{-1}. We have also used the G3(MP2) method,[28] which is known to be recommended for the energy calculation of fluorine compounds. The obtained relative energy value is –3.05 kJ mol^{-1}, virtually the same as our CCSD(T)//6-311++G(3df,3pd) value of –3.14 kJ mol^{-1}. Therefore, the ab initio post-HF level of calculation employed here can surely be considered very adequate.

At this point, from the analysis of the theoretical results reported in Tables 1−3 and Figures 1 and 2, it can be said that any disagreement with experimental results regarding the calculation of the anti/gauche conformational population ratio cannot be attributed only to the ab initio level of theory used in the present work. It can be seen that the variation in the calculated molecular properties (structural parameters, vibrational frequencies, and relative energies) as a function of the correlated ab initio level of theory and quality of the basis set would cause a change in the conformational population values, certainly well below the corresponding experimental uncertainties (2–5%). Therefore, we are confident in using our ab initio data to analyze the performance of the theoretical models for calculating thermal corrections through the evaluation of molecular partition functions, making use of the statistical thermodynamics formalism and, therefore, enthalpy and Gibbs free-energy values, leading to the theoretical determination of conformational population ratios.

Now, we finally turn to the analysis of the ab initio temperature-dependent enthalpy and Gibbs population results for the anti → gauche processes for 1,2-difluoroethane and 1,2-dichloroethane reported in Tables 4 and 5, with the temperature range being chosen according to the reported experimental conditions. Only the results obtained with the 6-311++G(3df,3pd) basis set and CCSD(T) level of theory are reported. Just to mention, as quoted in Tables 4 and 5, the MP2/aug-cc-pVTZ values for ΔU and ΔG_T show a very small deviation from the corresponding MP2/6-311++G-(3df,3pd) values (approximately 0.1 kJ mol^{-1} at 25 °C). Therefore, essentially the same conformational population is obtained using the aug-cc-pVTZ basis set.

It can be seen from Tables 4 and 5 that the effect of the anharmonic correction to the vibrational frequencies on the thermal energies, as explained in the computational section, is quite small (±0.04 kJ mol^{-1}) and so can be neglected; therefore, only the treatment of the low-frequency modes need to be considered. It is important to make it clear that the anharmonicity effect was not included explicitly in the vibrational partition function, which can easily been done for diatomic molecules;[18] however, much more work is required for polyatomic molecules. In the present case, the harmonic oscillator functional dependence was used for the vibrational partition function, but the anharmonic frequencies are utilized instead of harmonic values.

As far as enthalpy calculations are concerned, it can be seen from Table 4 that the ab initio and experimental enthalpy values for the anti → gauche process exhibit a very fair agreement, for both 1,2-dichloroethane and 1,2-difluorethane, independent of the way that the low-frequency modes are treated. In other words, the internal energy contribution is not so sensitive to the model used to treat the low-frequency modes in the calculation of relative enthalpy values, with the $\Delta E_{\text{ele-nuc}}$ contribution being of major relevance.

Table 4. Temperature-Dependent Enthalpy (ΔH) Values Calculated Including Anharmonicity and Hindered-Rotation Effects on the Internal Energy Correction (ΔU) Calculated at the MP2/6-311++G(3df,3pd) Level, for the anti → gauche Interconversion Process for 1,2-Dichloroethane and 1,2-Difluorethane[a]

1,2-dichloroethane	ΔU[b]	$\Delta U^{\text{True-Vib}}$[c]	$\Delta U^{\text{Hind-Rot}}$[d]	ΔH	$\Delta H^{\text{True-Vib}}$[c]	$\Delta H^{\text{Hind-Rot-Anh}}$[e]	ΔH exptl.
T = 25 °C	−0.50	−0.96	−0.04	4.98	4.52	4.90	[5.0 ± 0.8][f]
T = 40 °C	−0.50	−0.92	−0.04	4.98	4.52	4.90	
T = 140 °C	−0.54	−0.84	−0.04	4.94	4.64	4.85	

1,2-difluorethane	ΔU[b]	$\Delta U^{\text{True-Vib}}$[c]	$\Delta U^{\text{Hind-Rot}}$[d]	ΔH	$\Delta H^{\text{True-Vib}}$[c]	$\Delta H^{\text{Hind-Rot-Anh}}$[e]	ΔH exptl.
T = 25 °C	−0.63	−0.96	−0.08	−3.77	−4.10	−3.89	[−3.39 ± 0.54][g]
T = 56 °C	−0.67	−0.96	−0.13	−3.81	−4.10	−3.97	
T = 92 °C	−0.67	−0.96	−0.13	−3.81	−4.10	−3.97	

[a] CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) $\Delta E_{\text{ele-nuc}}$ values (5.48 and −3.14 kJ mol^{-1} for 1,2-dichloroethane and 1,2-difluorethane, respectively) were used. All values are in kJ mol^{-1}. [b] The MP2/aug-cc-pVTZ ΔU values for 1,2-dichloroethane and 1,2-difluorethane are respectively –0.50 and –0.63 kJ mol^{-1} at 25 °C. [c] Calculated using the vibrational partition function evaluated excluding the low-frequency normal vibrational modes (three modes at room temperature). Only the true vibrational modes that can be satisfactorily described as harmonic oscillators were considered. [d] Internal rotation correction to the MP2/6-311++G(3df,3pd) internal energy term (ΔU) value (one internal rotation was identified for all four species). [e] $\Delta H^{\text{Hind-Rot-Anh}} = \Delta E_{\text{ele-nuc}} + \Delta U + \Delta U^{\text{Hind-Rot}} + \Delta U^{\text{Anh}}$. Value obtained including the anharmonicity and hindered internal rotation correction to calculation of the internal energy correction. The anharmonic correction to internal energy (ΔU^{Anh}) is –0.04 kJ mol^{-1} for both 1,2-dichloro- and 1,2-difluorethane, evaluated at the MP2/6-311++G(2d,2p) level and room temperature. This should be our best enthalpy value. [f] Experimental value from ref 38. [g] Experimental value from ref 8.

Table 5. Temperature-Dependent Gibbs Population (%Pop.) and Relative Gibbs Free Energy (ΔG) Values Calculated Including Anharmonicity and Hindered-Rotation Effects on the Entropy Contribution ($T\Delta S$) to the Thermal Energy Correction (ΔG_T) Calculated at the MP2/6-311++G(3df,3pd) Level, for the anti → gauche Interconversion Process for 1,2-Dichloroethane and 1,2-Difluorethane[a]

1,2-dichloroethane	$T\Delta S$[b]	$T\Delta S^{\text{True-Vib}}$[c]	$T\Delta S^{\text{Hind-Rot}}$[d]	ΔG	%Pop. anti	$\Delta G^{\text{True-Vib}}$[c]	%Pop. anti	$\Delta G^{\text{Hind-Rot-Anh}}$[e]	%Pop. anti	%Pop. exptl. anti
T = 25 °C	−0.46	0.54	1.67	5.44	90%	3.97	83%	3.77	82%	[78 ± 5%][f]
T = 40 °C	−0.46	0.63	1.76	5.44	89%	3.93	82%	3.68	81%	[77.0 ± 1.7%][g]
T = 140 °C	−0.67	0.71	2.30	5.61	84%	3.93	76%	3.26	72%	[67.5 ± 2.2%][g]

1,2-difluorethane	$T\Delta S$[b]	$T\Delta S^{\text{True-Vib}}$[c]	$T\Delta S^{\text{Hind-Rot}}$[d]	ΔG	%Pop. anti	$\Delta G^{\text{True-Vib}}$[c]	%Pop. anti	$\Delta G^{\text{Hind-Rot-Anh}}$[e]	%Pop. anti	%Pop. exptl. anti
T = 25 °C	−0.71	0.21	−0.17	−3.05	23%	−4.31	15%	−2.93	23%	[37 ± 5%][h]
T = 56 °C	−0.84	0.25	−0.21	−2.97	25%	−4.35	17%	−2.85	26%	[41 ± 5%][h]
T = 92 °C	−0.92	0.25	−0.21	−2.89	28%	−4.35	19%	−2.76	29%	[44 ± 5%][h]

[a] CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) $\Delta E_{\text{ele-nuc}}$ values (5.48 and −3.14 kJ mol^{-1} for 1,2-dichloroethane and 1,2-difluorethane respectively) were used. All values are in kJ mol^{-1}. [b] The MP2/aug-cc-pVTZ $T\Delta S$ values for 1,2-dichloroethane and 1,2-difluorethane are respectively –0.46 and –0.84 kJ mol^{-1} at 25 °C. The room-temperature MP2/6-311++G(3df,3pd) rotational entropy ($T\Delta S^{\text{rot}}$) contributions are 0.46 and 0.21 kcal mol^{-1} for 1,2-dichloroethane and 1,2-difluorethane, respectively (identical to the MP2/aug-cc-pVTZ values). [c] Calculated using the vibrational partition function evaluated excluding the low-frequency normal vibrational modes (three modes at room temperature). [d] Internal rotation correction to the MP2/6-311++G(3df,3pd) entropy term ($T\Delta S$) value (one internal rotation was identified for all four species). [e] $\Delta G^{\text{Hind-Rot-Anh}} = \Delta E_{\text{ele-nuc}} + \Delta G_T + \Delta G_T^{\text{Hind-Rot}} + \Delta G_T^{\text{Anh}}$; ($\Delta G_T = \Delta U - T\Delta S$). Values obtained include the anharmonicity and hindered internal rotation correction to the calculation of the thermal energy correction (ΔG_T). The anharmonic correction to entropy ($T\Delta S^{\text{Anha}}$) is –0.08 kJ mol^{-1}, for both 1,2-dichloro- and 1,2-difluorethane, evaluated at the MP2/6-311++G(2d,2p) level and room temperature. This should be our best Gibbs free-energy value. [f] Experimental value from ref 4. See also ref 38. [g] Experimental value from ref 40 [h] Experimental value from ref 8. There are two other population datapoints obtained from an electron diffraction experiment that differ considerably from the more recent reported value in ref 8 based on the vibrational spectroscopy analysis: 9% of the anti form from ref 41 at room temperature and 4.0 ± 1.8 at 22 °C from ref 42.

Regarding 1,2-dichloroethane, Ayala and Schlegel[19] have shown that the hindered-rotor approach would be appropriate in the evaluation of the vibrational partition function using the HF/6-31G(d) level of calculation,[39] which is also corroborated by the results reported here. In the present article, comparisons with experimental results are made for conformational populations, which were determined experimentally with a satisfactory precision, and calculated at the ab initio level using a specific vibrational partition function containing a treatment of the low-frequency modes reported by Ayala and Schlegel[19] (implemented in the Gaussian package[33]). Then, when the agreement between theoretical and experimental populations is analyzed, an assessment of the performance of the hindered-rotor approach can be made. From Table 5, the effectiveness of the hindered-rotor approach to describe the 1,2-dichloroethane species is promptly seen, leading to a good agreement with gas-phase electron diffraction conformational population data. The simple procedure of neglecting the low-frequency modes (three modes at room temperature) in the evaluation of the vibrational partition function, which may be considered as a rough but simple approximation, used successfully in the conformational analysis of cyclooctane[43,44] and cycloheptane,[45] also works well for 1,2-dichloroethane up to a temperature of 40 °C.

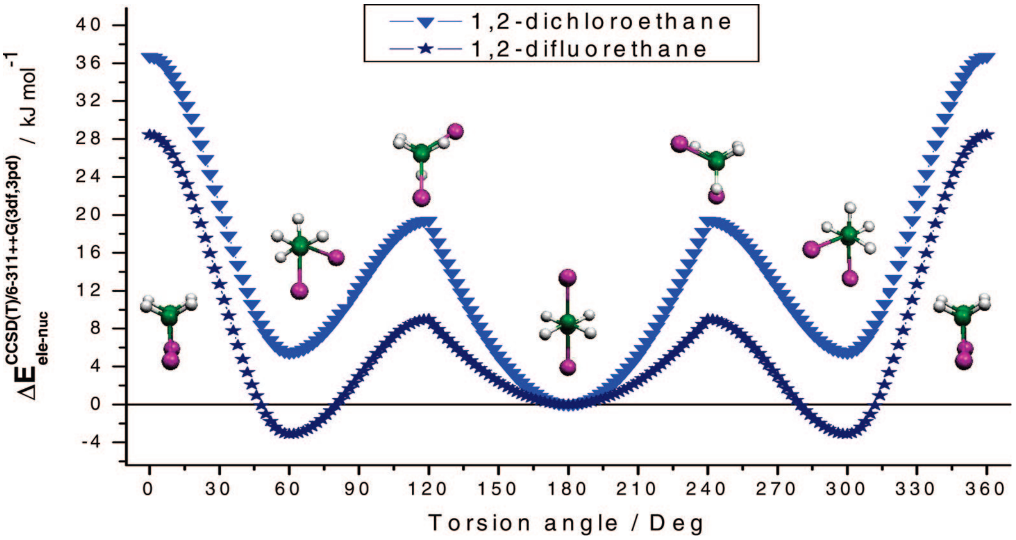


Figure 3. CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) relative energies (in units of kJ mol^{-1}) for the four stationary points located on the MP2/6-311++G(3df,3pd) PES for 1,2-dichlorothane and 1,2-difluorethane: **anti** minimum, **TS1** structure, **gauche** minimum, **TS2** structure. The room temperature V_0/kT barrier values are 32.6 kJ mol^{-1} (anti → gauche process) and 20.5 kJ mol^{-1} (gauche → anti process), respectively, for the chlorine and fluorine species (1 cal = 4.184 J; 1 kcal mol^{-1} = 349.38 cm^{-1}).

For 1,2-difluorethane, a satisfactory agreement with experimental conformational population data was not obtained. An interesting feature that can be seen from Table 5 is the fact that the procedure of treating the lowest-frequency modes as a hindered rotor leads to a very small correction, compared to the corresponding value obtained for 1,2-dichloroethane, providing virtually the same conformational population as the consideration of all 3N-6 modes as harmonic oscillators. So, in this case, the procedure was useless. The alternative of ignoring the three lowest-frequency modes also does not work well here, similar to the case of the cyclononane molecule.[46]

The calculated CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) potential barriers (V_0) are shown in Figure 3 for the anti ↔ TS1 ↔ gauche and gauche ↔ TS2 ↔ gauche processes. It can be seen from Figure 3 that the energy barrier for anti → gauche interconversion is 19.2 (13.8) kJ mol^{-1} for 1,2-dichloroethane and 12.1 (9.2) kJ mol^{-1} for the gauche → anti interconversion in 1,2-difluorethane (the values in parenthesis are for the reverse process). The corresponding barriers for the interconversion between two equivalent gauche structures are 31.4 and 31.8 kJ mol^{-1} respectively for 1,2-dichloroethane and 1,2-difluorethane. The gauche → anti barrier for the fluorine species is 7.1 kJ mol^{-1} (4.6 kJ mol^{-1} for the reverse barrier) smaller than the anti → gauche barrier for 1,2-dichloroethane. The experimental barriers reported for 1,2-difluorethane[8] are 8.87 (anti ↔ TS1 → gauche) and 23.93 kJ mol^{-1} (gauche → TS2 → gauche). The former is close to the CCSD(T)/6-311++G(3df,3pd)//MP2/6-311++G(3df,3pd) calculated value (12.1 kJ mol^{-1}), whereas the latter is predicted to be 31.8 kJ mol^{-1}, a deviation of almost 8 kJ mol^{-1}. At temperatures such that $V_0 \ll kT$ (where k is the Boltzmann constant and T the absolute temperature), the internal rotation is essentially free and can be treated by methods similar to those for the rigid rotor, and when $V_0 \gg kT$, the molecule is trapped at the bottom of the potential well and the motion is that of a simple torsional vibrational, which can be treated by a method similar to that used for the simple harmonic oscillator. For intermediate V_0 values, as the ones reported in this work for 1,2-difluorethane (V_0/kT = 20.5 kJ mol^{-1}) and 1,2-dichloroethane (V_0/kT = 32.6 kJ mol^{-1}), at room temperature, the motion is intermediate between that of a free rotation and that of torsional vibration. In this work, we used a modified hindered rotor partition function for the lowest-frequency vibrational mode proposed by Ayala and Schlegel,[19] which has a dependence on V_0/kT, as previously reported by Pitzer and Gwinn.[29] We assumed that the molecular vibrational partition function (q^{vib}) can be written as a product of harmonic oscillator (HO) and hindered rotor (Hind-Rot) contributions according to eq 6, and so the thermodynamic functions are obtained as a sum of two terms, and then the hindered rotor correction is analyzed.

$$q^{\text{vib}} = q^{\text{HO}} q^{\text{Hind-Rot}} \qquad (6)$$

Looking at the individual values for the hindered rotor correction, it can be seen that the problem is with the gauche form of 1,2-difluorethane. For 1,2-dichloroethane, the values of the respective corrections for entropy contribution are $\text{TS}_{\text{anti}}^{\text{Hind-Rot}}$ = 0.17 and $\text{TS}_{\text{gauche}}^{\text{Hind-Rot}}$ = 1.84 kJ mol^{-1}. However, for 1,2-difluorethane, the values are rather different, with the correction for the gauche structure being even less than that for the anti form, that is, $\text{TS}_{\text{anti}}^{\text{Hind-Rot}}$ = 0.33 and $\text{TS}_{\text{gauche}}^{\text{Hind-Rot}}$ = 0.17 kJ mol^{-1}. Therefore, it appears that the hindered rotor correction for the 1,2-difluorethane gauche conformer should be similar to that for the 1,2-dichloroethane (around 1.67 kJ mol^{-1}), however, having a negative value. We cannot say for sure if only the variation in V_0 with respect to 1,2-dichloroethane affects so much the hindered rotor partition function given in ref 19 to the point of making the agreement with experimental conformational population so poor.

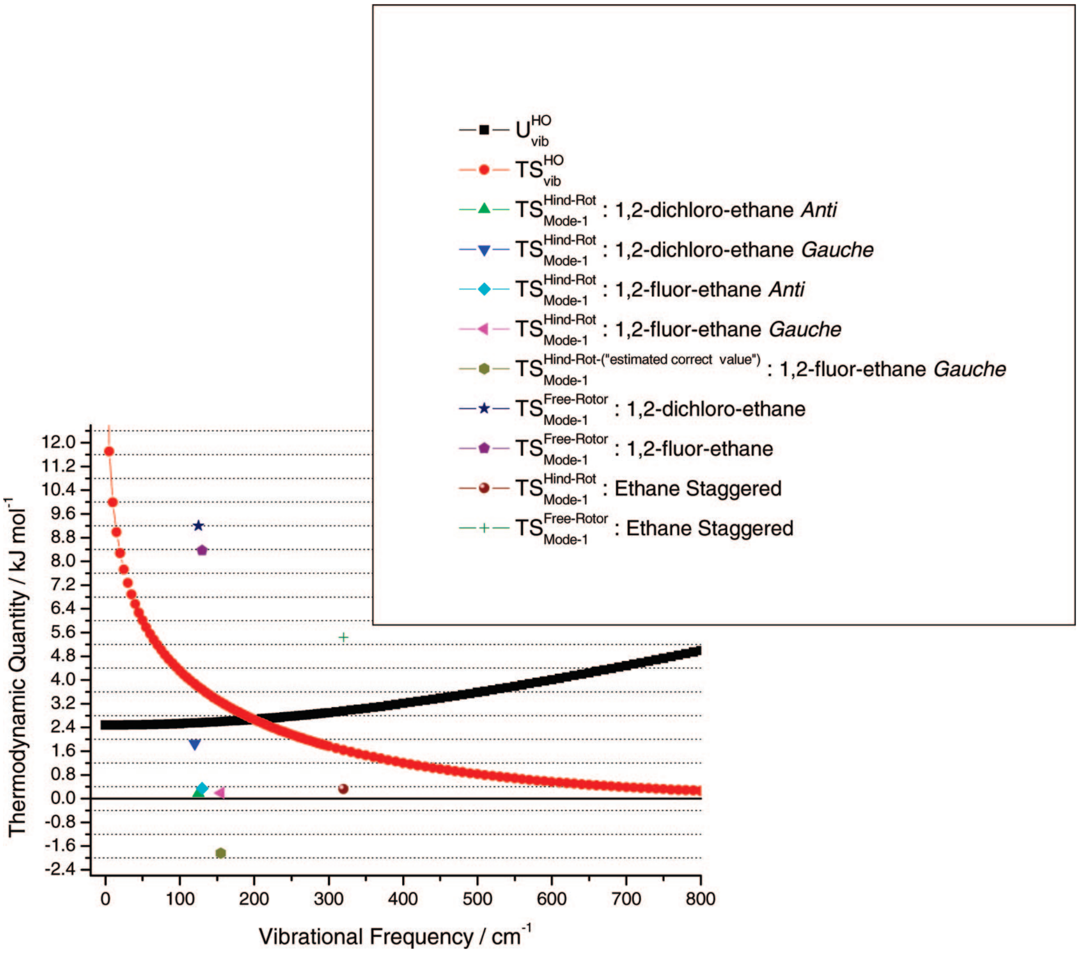


Figure 4. Thermodynamic energy or internal thermal energy (**U**) and entropic (**TS**) vibrational contributions (in units of kJ mol^{-1}) represented as a function of the vibrational frequency, calculated with the aid of the statistical thermodynamics formulas, within the harmonic oscillator (HO) approximation (HO vibrational partition function), at room temperature and normal pressure. The specific TS vibrational contribution due to the lowest frequency mode, using a hindered-rotor partition function and free-rotor approach is also shown for ethane, 1,2-chloro- and 1,2-difluoroethane (MP2/6-311++G(3df,3pd) value).

With the aim of clarifying the reason for the disagreement between theoretical and experimental gas-phase conformational population for 1,2-difluorethane, we decided to use the experimental entropy for the anti ↔ gauche process, obtained from the analysis of the vibrational spectral data dependence with temperature reported in ref 8, where by applying the van't Hoff isochore equation the entropy change for the process could be evaluated. The experimental entropy contribution at room temperature is $T\Delta S^{Expt}$ = –2.05 kJ mol^{-1}. Our MP2/6-311++G(3df,3pd) best value is –0.71 kJ mol^{-1} (a quite sizable 65% difference). Using the experimental entropy and our ab initio CCSD(T)/6-311++G-(3df,3pd) relative energy ($\Delta E_{ele-nuc}$) and MP2/6-311++G-(3df,3pd) internal energy (ΔU), we obtain a room-temperature Gibbs population of 33% of the anti form, in good agreement with the experimental value of 37 ± 5%. Therefore, it is quite evident that our calculated entropy for the anti ↔ gauche process of 1,2-difluorethane, using the combined quantum mechanical/statistical thermodynamic approach, is in serious error. It is well-known that the entropy term ($T\Delta S^{vib}$) has a much higher sensibility to the low-frequency mode than the internal energy (ΔU^{vib}), which can be easily seen from Figure 4, where the variation of the respective thermodynamic functions with the vibrational frequency is shown. ΔU^{vib} is very monotonically dependent on the frequency in the low-frequency region, which explains why our calculated enthalpies are in good agreement with the experimental ones. On the contrary, the entropy counterpart is strongly dependent on the frequency in the region of 0–200 cm^{-1}; therefore, the treatment of low-frequency modes definitively has a pronounced effect on the entropy evaluation. When we used the experimental entropy for the gauche → anti process (1,2-difluorethane) instead of the calculated one using the hindered rotor approach, the agreement with experimental results is fine. This is irrefutable proof that the calculated $T\Delta S^{vib}$ entropy contribution to the Gibbs free energy (ΔG) is poorly described, and it becomes evident that this is the major reason for the serious disagreement with the experimental determination of the conformational population for the 1,2-difluorethane. Also shown in Figure 4 is the individual hindered-rotor corrections for each conformer (anti and gauche) and also for the ethane molecule only, for reasons of comparison.

The calculated ab initio (MP2/6-311++G(3df,3pd) value) TS^{vib} contribution due to the first low-frequency mode for the individual anti and gauche forms of 1,2-dichloro- and

1,2-difluorethane, and also the ethane molecule (*staggered* conformation), evaluated using the hindered-rotor approach from ref 19 (named $TS^{Hind-Rot}$) and also considering the first vibrational mode as a free rotor ($TS^{Free-Rotor}$) are highlighted in Figure 4 (the free-rotor contribution is the same for the anti and gauche forms, since there is no barrier for rotation, so the effect on the entropy difference is null, which is equivalent to just ignoring the first low-frequency mode in the evaluation of the vibrational partition function). The hindered-rotor correction for the anti form of the fluor and chlorine species is almost the same. However, for the gauche form, the correction for the chlorine is quite different from the fluorine species, which practically has no variation at all, as already mentioned before. Then, the contribution of the hindered rotation to entropy is virtually zero for the 1,2-difluorethane, and this can be a reason for the disagreement with experimental results. Analyzing the rotational constants and spatial orientations with respect to the principal axis coordinate system, and using the rotational entropy data for CH_3Cl and CH_3F and the corresponding values for $^{\bullet}CH_2Cl$ and $^{\bullet}CH_2F$ radicals as a model for a free rotating group, assuming the first low-frequency mode as a free rotor, we estimated the contribution to the vibrational entropy ($TS^{Free-Rotor}$) to be ~9.2 and 8.4 kJ mol^{-1} respectively for the chlorine and fluorine species. In the case of the ethane molecule, when the rotational entropy data for CH_4 and the $^{\bullet}CH_3$ radical is used for modeling the free rotating CH_2 group, the free rotor value for the first vibrational mode is estimated to be ~5.4 kJ mol^{-1}. Therefore, there is no point in considering the internal rotation mode as a free rotor, in the case of 1,2-difluorethane. The sizable difference in the energy barrier for internal rotation between the chlorine ($V_0 = 19.2$ kJ mol^{-1}) and fluorine ($V_0 = 12.1$ kJ mol^{-1}) species (anti → gauche and gauche → anti processes respectively) would probably make a contribution to the poor performance of the hindered rotor partition function in the case of 1,2-difluorethane, and consequently a visible disagreement with experiment, regarding the anti/gauche conformational population ratio values. However, we cannot quantify the extent of this effect. We may just speculate that the vibrational partition used in this work may have an inadequate dependence on the V_0/kT value, which could make it unsuitable for substituted ethane molecules exhibiting the gauche effect, such as 1,2-difluorethane. If only the lowest-frequency mode is ignored, essentially the same result is obtained when it is considered a harmonic oscillator in the vibrational partition function for 1,2-dichloroethane, corroborating that it should be treated separately as a hindered rotor, and not just disregarded, where, as can be seen in Figure 4, it adds a substantial correction to the anti/gauche entropy difference ($T\Delta S^{Hind-Rot} = 1.67$ kJ mol^{-1}) and therefore conformational population. Nevertheless, the same reasoning does not apply to 1,2-difluorethane, where the treatment of the first vibrational normal mode as a hindered rotor did not make any noticeable difference ($T\Delta S^{Hind-Rot} = –0.17$ kJ mol^{-1}). For the ethane molecule, the effect due to the hindered rotation is also small ($TS^{Hind-Rot} = 0.31$ kJ mol^{-1}; $S^{Hind-Rot} = 1.05$ J mol^{-1} K^{-1}), but improving the deviation from experimental results from 1% to 0.5%. So, in this well-known case, the hindered-rotor model works perfectly well.

It is also opportune to emphasize here that, as already pointed out by Ayala and Schlegel,[19] in principle, most of the problem resides in the identification of the internal rotation modes. Large molecules can have a large number of low-frequency modes which can include not only internal rotations but also large amplitude collective bending motions of atoms. Moreover, some of the low-frequency modes can be a mixture of such motions. In the case of cyclic molecules featuring rings bigger than six members in size, there are ring torsional modes, and similar to internal rotations, ring torsions can cause problems in the evaluation of thermodynamic functions (see for example refs 44–46). For simple systems such as substituted ethane molecules, it is possible to unambiguously identify the low-frequency internal rotation modes and also use an adequate ab initio correlated level of theory, with a good quality basis set, to provide equilibrium geometries and vibrational frequencies necessary for the evaluation of partition functions, and also relative energy values, for the calculation of thermodynamic quantities. In this case, an assessment of the performance of approximated methods available to predict conformational populations can be made.

4. Conclusions

It is usually assumed that the standard statistical thermodynamics formalism can be safely applied once the quality of the distinct energy values, which are necessary for the evaluation of partition functions, is assured by the use of accurate quantum chemical methods. In this work, we used the best computational affordable quantum chemical methods in the calculation of geometrical parameters and harmonic frequencies (MP2/6-311++G(3df,3pd) and MP2/aug-cc-pVTZ), required for the evaluation of rotational and vibrational partition functions (and therefore thermal correction, ΔG_T), and highly correlated post-HF calculations (CCSD(T)/6-311++G(3df,3pd) and CCSD(T)/aug-cc-pVQZ) of relative energies ($\Delta E_{ele-nuc}$), required for the theoretical calculation of Gibbs free energies (ΔG) and, therefore, conformational population values. We selected two substituted ethane molecules, where for one of them (1,2-difluorethane) the gauche effect is known to be operating. In the calculation of thermal corrections, we used a hinder-rotor treatment for the low-frequency mode and also anharmonicity correction to the vibrational frequencies within the harmonic oscillator partition function. The agreement with gas-phase experimental data was fine for 1,2-dichloroethane; however, reasonable accordance with experimental results was not obtained in the case of 1,2-difluorethane. The cause of the disagreement with the experimental conformational population ratio is not the ab initio level of calculation employed, since we analyzed the behavior of the methodology as a function of level of theory and size of the basis set and could guarantee that, by improving the level of calculation to a computational unreachable degree of sophistication, no significant variation of the conformational population would be observed. Therefore, the results reported here show definitively that the problem of the agreement with experiment is related to the treatment used for the low-frequency vibrational modes. The inclusion of anharmonic corrections (using anharmonic frequencies and the harmonic oscillator partition function) and a treatment of low-frequency modes as a hindered rotor for the evaluation of the thermal correction ($\Delta G_T^{\text{Hind-Rot-Anh}}$) did not improve the agreement with experimental results regarding conformational population values, in the case of 1,2-difluorethane. The lower gauche → anti interconversion energy barrier found for 1,2-difluorethane, compared to that for 1,2-dichloroethane (anti → gauche), may contribute to the poor performance of the hindered rotor partition function; however, we cannot quantify its extent nor how important the gauche effect is for the whole affair.

It is interesting to point out that a great part of the molecular systems of interest to a number of quantum chemists, using standard computer codes (freely delivered or not) where the quantum mechanical and statistical thermodynamics formalism are readily implemented in the calculation of thermodynamic properties, may have peculiarities that can easily be overlooked. In this case of substituted ethane molecules, only three low-frequency modes are present, and therefore, a theoretical, sound, and adequate computational model is available or can be pursued. However, for large molecular systems, a considerable number of low-frequency modes are certainly present, and use of the harmonic oscillator partition function is not advised. So, in this case, finding a satisfactory computational treatment is indeed a hard task. Fortunately, in many situations, energy differences are calculated, and so, a cancelation of errors is often present. Then, the effect of the low-frequency modes may be of small importance, mainly when the size of the relative energy value ($\Delta E_{\text{ele-nuc}}$) is much larger than the thermal energy correction.

Acknowledgment. The authors would like to thank the Brazilian agencies CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) for financial support. This work is part of the project PRONEX-FAPEMIG/EDT-537/05.

References

(1) Roberts, J. D. Fascination with the Conformational Analysis of Succinic Acid, as Evaluated by NMR Spectroscopy, and Why. *Acc. Chem. Res.* **2006**, *39*, 889. and references therein.

(2) Freitas, M. P.; Rittner, R. Is There a General Rule for the Gauche Effect in the Conformational Isomerism of 1,2-disubstituted Ethanes? *J. Phys. Chem. A* **2007**, *111*, 7233. and references therein.

(3) Orville-Thomas, W. J. *Internal Rotation in Molecules*; John Wiley & Sons: London, 1974; pp 101.

(4) Ainsworth, J.; Karle, J. The Structure and Internal Motion of 1,2-Dichloroethane. *J. Chem. Phys.* **1952**, *20*, 425.

(5) Youssoufi, Y. E.; Herman, M.; Lievin, J. The ground electronic state of 1,2-dichloroethane I. Ab initio investigation of the geometrical, vibrational and torsional structure. *Mol. Phys.* **1998**, *94*, 461.

(6) Wiberg, K. B.; Murcko, M. A. Rotational Barriers. 1. 1,2-Dihaloethanes. *J. Phys. Chem.* **1987**, *91*, 3616.

(7) Hirano, T.; Nonoyama, S.; Miyajima, T.; Kurita, Y.; Kawamura, T.; Sato, H. Gas-phase and ^{1}H High-resolution N.M.R. Spectroscopy: Application to the Study of Unperturbed Conformational Energies of 1,2-Difluoroethane. *J. Chem. Soc., Chem. Commun.* **1986**, 606.

(8) Durig, J. R.; Liu, J.; Little, T. S.; Kalasinsky, V. F. Conformational Analysls, Barriers to Internal Rotation, Vibrational Assignment, and ab Initio Calculations of 1,2-Difluoroethane. *J. Phys. Chem.* **1992**, *96*, 8224.

(9) Pophristic, V.; Goodman, L. Hyperconjugation not steric repulsion leads to the staggered structure of ethane. *Nature* **2001**, *411*, 565.

(10) Bickelhaupt, F. M.; Baerends, E. J. The Case for Steric Repulsion Causing the Staggered Conformation of Ethane. *Angew. Chem., Int. Ed.* **2003**, *42*, 4183.

(11) Goodman, L.; Gu, H.; Pophristic, V. Gauche Effect in 1,2-Difluoroethane. Hyperconjugation, Bent Bonds, Steric Repulsion. *J. Phys. Chem. A* **2005**, *109*, 1223.

(12) Goodman, L.; Sauers, R. R. 1-Fluoropropane. Torsional Potential Surface. *J. Chem. Theory Comput.* **2005**, *1*, 1185.

(13) Eliel, E. L.; Wilen, S. H.; Mander, L. N. Conformation of Acyclic Molecules. In *Stereochemistry of Organic Compounds*; John Wiley & Sons Inc., Wiley-Interscience Publication: New York, 1994; pp 599.

(14) Travis, K. P.; Searles, D. J. Effect of solvation and confinement on the *trans-gauche* isomerization reaction in *n*-butane. *J. Chem. Phys.* **2006**, *125*, 164501.

(15) Brunck, T. K.; Weinhold, F. Quantum-Mechanical Studies on the Origin of Barriers to Internal Rotation about Single Bonds. *J. Am. Chem. Soc.* **1979**, *101*, 1700.

(16) Kemp, J. D.; Pitzer, K. S. Hindered Rotation of the Methyl Groups in Ethane. *J. Chem. Phys.* **1936**, *4*, 749.

(17) Pitzer, R. M. The Barrier to Internal Rotation in Ethane. *Acc. Chem. Res.* **1983**, *16*, 207. and references therein.

(18) See for example: McQuarrie, D. A. *Statistical Thermodynamics*; University Science Books: Mill Valley, CA, 1973; pp 129–141.

(19) Ayala, P. Y.; Schlegel, H. B. Identification and treatment of internal rotation in normal mode vibrational analysis. *J. Chem. Phys.* **1998**, *108*, 2314. and references therein.

(20) Ellingson, B. A.; Lynch, V. A.; Mielke, S. L.; Truhlar, D. G. Statistical thermodynamics of bond torsional modes: Tests of separable, almost-separable, and improved Pitzer-Gwinn approximations. *J. Chem. Phys.* **2006**, *125*, 084305. and references therein.

(21) Barone, V. Vibrational zero-point energies and thermodynamic functions beyond the harmonic approximation. *J. Chem. Phys.* **2004**, *120*, 3059. and references therein.

(22) Barone, V. Anharmonic vibrational properties by a fully automated second-order perturbative approach. *J. Chem. Phys.* **2005**, *122*, 014108.

(23) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618.

(24) (a) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3–21+G basis set for first-row elements, Li-F. *J. Comput. Chem.* **1983**, *4*, 294. (b) Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular

orbital methods 25. Supplementary functions for Gaussian basis sets. *J. Chem. Phys.* **1984**, *80*, 3265.

(25) (a) Woon, D. E.; Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **1993**, *98*, 1358. (b) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796.

(26) See for example: Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry, Introduction to Advanced Electronic Structure Theory*; Dover Plublications, Inc.: New York, 1996; pp 320–379.

(27) Bartlett, R. J.; Stanton, J. F. In *Reviews In Computational Chemistry, Aplications of Post-Hartree-Fock Methods*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers, Inc: New York, 1994; *Vol. 5*, Chapter 2, pp 65–169.

(28) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. Gaussian-3 theory using reduced Møller-Plesset order. *J. Chem. Phys.* **1999**, *110*, 4703. and references therein.

(29) Pitzer, K. S.; Gwinn, W. D. Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation I. Rigid Frame with Attached Tops. *J. Chem. Phys.* **1942**, *10*, 428.

(30) Li, J. C. M.; Pitzer, K. S. Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation. IV. Extended Tables for Molecules with Small Moments of Inertia. *J. Phys. Chem.* **1956**, *60*, 466.

(31) Truhlar, D. G. A Simple Approximation for the Vibrational Partition Function of a Hindered Internal Rotation. *J. Comput. Chem.* **1991**, *12*, 266.

(32) McClurg, R. B.; Flagan, R. C.; Goddard, W. A. The hindered rotor density-of-states interpolation function. *J. Chem. Phys.* **1997**, *106*, 6675.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, J. M. C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.04; Gaussian, Inc.: Pittsburgh, PA, 2003.

(34) See also: Hirota, E.; Saito, S.; Endo, Y. Barrier to internal rotation in ethane from the microwave spectrum of CH_3CHD_2. *J. Chem. Phys.* **1979**, *71*, 1183.

(35) Kemp, J. D.; Pitzer, K. S. The Entropy and the Third Law of Thermodynamics. Hindered Rotation of Methyl Groups. *J. Am. Chem. Soc.* **1937**, *59*, 276.

(36) Takeo, H.; Matsumura, C.; Morino, Y. Microwave spectrum and molecular structure of gauche-1,2-difluoroethane. *J. Chem. Phys.* **1986**, *84*, 4205.

(37) Xavier, E. S.; Rocha, W. R.; Da Silva, J. C. S.; Dos Santos, H. F.; De Almeida, W. B. *Ab initio* thermodynamic study of the reaction of CF_2Cl_2 and CHF_2Cl CFCs species with OH radical. *Chem. Phys. Lett.* **2007**, *448*, 164.

(38) Bernstein, H. J. Internal Rotation II. The Energy Difference between the Rotational Isomers of 1,2-Dichloroethane. *J. Chem. Phys.* **1949**, *17*, 258.

(39) Chung-Phillips, A. Torsional energy levels and wave functions. *J. Comput. Chem.* **1992**, *13*, 874.

(40) Kveseth, K. Conformational Analysis. 1. The Temperature Effect on the Structure and Composition of the Rotational Conformers of 1,2-Dichloroethane as Studied by Gas Electron Diffraction. *Acta Chem. Scand. A* **1974**, *28*, 482.

(41) Fernholt, L.; Kveseth, K. Conformational-Analysis. The Temperature Effect on the Structure and Composition of the Rotational Conformers of 1,2-Dichloroethane as Studied by Gas Electron-Diffraction. *Acta Chem. Scand. A* **1980**, *34*, 163.

(42) Friesen, D.; Hedberg, K. Conformational Analysis. 7. 1,2-Difluoroethanee An Electron-Diffraction Investigation of the Molecular Structure, Composition, Trans-Gauche Energy and Entropy Differences, and Potential Hindering Internal Rotation. *J. Am. Chem. Soc.* **1980**, *102*, 3987.

(43) De Almeida, W. B. Molecular structure determination of cyclooctane by ab initio and electron diffraction methods in the gas phase. *Quim. Nova* **2000**, *23*, 600.

(44) Dos Santos, H. F.; Rocha, W. R.; De Almeida, W. B. On the evaluation of thermal corrections to gas phase ab initio relative energies: implications to the conformational analysis study of cyclooctane. *Chem. Phys.* **2002**, *280*, 31.

(45) Anconi, C. P. A.; Nascimento, C. S., Jr.; Dos Santos, H. F.; De Almeida, W. B. A highly correlated ab initio investigaton of the temperature-dependent conformational analysis of cycloheptane. *Chem. Phys. Lett.* **2006**, *418*, 459.

(46) Franco, M. L.; Ferreira, D. E. C.; Dos Santos, H. F.; De Almeida, W. B. Temperature-dependent conformational analysis of cyclononane: An ab initio study. *Int. J. Quantum Chem.* **2007**, *107*, 545.

CT800014W

JCTC Journal of Chemical Theory and Computation

Structural and Energetic Study of Cisplatin and Derivatives: Comparison of the Performance of Density Funtional Theory Implementations

Pablo D. Dans,[†] Alejandro Crespo,[‡] Darío A. Estrin,[‡] and E. Laura Coitiño*,[†]

Laboratorio de Química Teórica y Computacional, Instituto de Química Biológica, Facultad de Ciencias, Universidad de la República (UdelaR), Centro Universitario Malvín Norte, Iguá 4225, Montevideo 11400, Uruguay, and Departamento de Química Inorgánica, Analítica y Química Física/IUIMAE-CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón II, Buenos Aires (C1428EHA), Argentina

Received September 18, 2007

Abstract: In this work, we compare the performance of different DFT implementations, using analytical and numerical basis sets for the expansion of the atomic wave function, in determining structural and energetic parameters of Cisplatin and some biorelevant derivatives. Characterization of the platinum-containing species was achieved at the HF, MP2, and DFT (PBE1PBE, mPW1PW91, B3LYP, B3PW91, and B3P86) levels of theory, using two relativistic effective core potentials to treat the Pt atom (LanL2DZ and SBK), together with analytical Gaussian-type basis sets as implemented in Gaussian03. These results were compared with those obtained with the SIESTA code that employs a pseudopotential derived from the Troullier−Martins procedure for the Pt atom and numerical pseudoatomic orbitals as basis set. All modeled properties were also compared with the experimental values when available or to the best theoretical calculations known to date. On the basis of the results, SIESTA is an excellent alternative to determine structure and energetics of platinum complexes derived from Cisplatin, with less computational efforts. This validates the use of the SIESTA code for this type of chemical systems and thus provides a computationally efficient quantum method (capable to linear scaling at large sizes and available in QM/MM implementations) for exploring larger and more complex chemical models which shall reproduce more faithfully the real chemistry of Cisplatin in physiological conditions.

Introduction

Cisplatin (cis-diamminedichloroplatinum(II)) is one of the most used drugs against cancer, being particularly effective in the treatment of testis, ovary, head and neck, bladder, and lung malignancies.[1] Despite the research efforts accumulated during the last 30 years addressed to elucidate the mode of action of the drug at the cellular and biochemical levels,[1,2a–e] the knowledge of the intimate chemical interactions established by the drug with relevant biomolecules (which determine cellular sensitiveness and resistance) is still at a stage far from satisfactory.[2d,e]

Reaching a deep understanding of these issues both in the case of Cisplatin and other active analogues requires not only a detailed characterization of the molecular mechanism of aquation (successive substitution of the labile ligands by water molecules, a process nowadays recognized as the activation step of these drugs in the cell)[2d] but also a complete study of the interaction and covalent binding of the drug to DNA, their pharmacological target. In the last

* To whom correspondence should be addressed. E-mail: lcoitino@luna.fcien.edu.uy. Fax: (598-2) 525 0749.

[†] Universidad de la República (UdelaR), Centro Universitario Malvín Norte.

[‡] Universidad de Buenos Aires, Ciudad Universitaria.

10.1021/ct7002385 CCC: $40.75
Published on Web 04/19/2008

two decades, theoretical and computational chemistry has started to make significant contributions on these and other related topics,[3–6] essentially by means of studies using reduced representations of the corresponding biological systems. Whereas gaining insight into the details of reactivity, interactions, and chemical processes established between platinum drugs and water/DNA requires the use of first principles quantum correlated methods—density functional theory (DFT) being the current privileged choice—size and complexity of the systems under physiological conditions turns necessary the use of hybrid quantum classical (QM/MM) descriptions to remain the study affordable. An efficient computational package available for this kind of combined quantum/classical descriptions is based upon an implementation of DFT using numerical basis sets.[7] For that reason, in this work we will validate technical choices as well as to ascertain the reliability of the outcomes of modeling regarding structure, reactivity and energetics of platinum-containing systems.

Although more than 30 articles[3–6] have appeared since the early 1980s applying quantum chemistry methods to characterize molecular properties (geometrical, electronic and vibrational, both as a goal by themselves[3a,f–i,4a,b] or aimed to develop force-field parameters[3b,c]) of Cisplatin and related compounds, participant species, thermodynamics and kinetics of aquation processes of platinum square planar complexes,[4d,5] and platination of DNA nucleobases,[4c–f,6] most of them employed implementations of the correlated theoretical level of choice—DFT within them—using different schemes of relativistic effective core potentials (ECPs) together with Gaussian-type basis sets, the use of DFT implementations with other kinds of basis sets being essentially limited to few cases addressing the most complex studies (both using plane waves[4c–f] or numerical-atomic orbitals as basis set[3f,5a,d,n,6j,k]). Only a marginal part of this work done during almost 25 years has been addressed to systematically assess the performance at the task of different quantum mechanics methods[3d,e,5b,e] (including several different DFT exchange/correlation functionals and ECPs/basis set schemes) and the relevance of including well balanced solvent/environment representations[4d–f,5b,h,j–n] in coping with some of the aforementioned issues.

A first key article pioneering in validating the use of DFT (BLYP) to describe the structure and bonding of Cisplatin, Transplatin, and their mono/diaqua-substituted derivatives was published in 1995 by Carloni et al.[4a] using plane waves for the expansion of the electronic wave function. Geometrical parameters and vibrational frequencies in fairly good agreement with the observed values were obtained, and frontier Kohn–Sham orbitals were also analyzed therein, finding for the first time a highest occupied molecular orbital−lowest unoccupied molecular orbital (HOMO–LUMO) gap consistent with experiment. Some years later, Pavankumar et al.[3d] provided a comprehensive test on the performance of HF and Möller–Plesset levels in modeling Cisplatin properties mostly using a wide range of Pople's basis sets and different ECPs schemes. More recently, whereas Wysokinski and Michalska[3e] analyzed the behavior of several pure and hybrid DFT functionals (using different ECP/basis set combinations) in predicting structural parameters, bonding and IR frequencies of Cisplatin and Carboplatin. Zhang et al. [5e] pursued a similar study focusing on the geometrical structure of Cisplatin, $PtCl_4^{2-}$, and water. Both groups agreed in selecting mPW1PW91 as the most reliable functional. However, none of these studies systematically addressed neither the performance in predicting energetics of the aquation processes (a task whose results reported by different groups have raised some controversies, their quality being sensitive to the ECP/basis set employed[5e] as well as to the description of microscopic/bulk solvent effects[5b,h,j–n]) nor that in describing bonding, thermodynamics and kinetics of the reaction between Cisplatin and DNA, even resorting to minimal models of the latter (bare nucleobases/nucleotides).

Evaluation of the existing DFT functionals and implementations, in particular those extensively used at the present for studying Cisplatin chemistry (mostly hybrid functionals), becomes necessary to calibrate and extend current knowledge to larger molecular systems and helps in the development of better methodological tools. On the other hand, the need to answer several questions in our own applied research regarding the mechanism of action of Cisplatin and related species in complex biological environments (i.e., including different double strand B-DNA sequences, Na^+ counterions and a realm of water molecules)[8] using proper structures and energetics, obtained with affordable strategies such as those present in the SIESTA (Spanish Initiative for Electronic Simulations with Thousands of Atoms) package,[7] was another practical motivation for pursuing the study presented here.

In terms of computational cost, construction of the electronic wave functions based on numerical-atomic orbitals has a great advantage over analytical localized basis sets of comparable accuracy. This approximation is used by DFT implementations in this code, providing computationally efficient quantum methods capable to linear scaling at large sizes (>100 atoms), also very efficient at intermediate molecular sizes. SIESTA uses the standard Kohn–Sham self-consistent density functional method both in the local density (LDA-LSD) and generalized gradient (GGA) approximations with norm-conserving pseudopotentials.[7] This type of quantum mechanics software can be used for the study of medium size or large systems respectively using QM (i.e., structure of the drugs and simplified representations of their chemical transformations) or QM/MM strategies (i.e., covalent interactions between water/DNA and Cisplatin under physiological conditions). To do so, validation of key issues coming from modeling against experimental data becomes mandatory, as well as comparing the performance of these implementations using numerical basis sets against other standard implementations of DFT using analytical ones, such as those currently included in GAUSSIAN[9a] or GAMESS,[9b] the main computational tools chosen in pursuing recent studies of Cisplatin and other platinated drugs interactions.[3g–i,5g–m,6q,r]

The model situations chosen in the present work to validate and compare different DFT strategies are the following: (i) Cisplatin; (ii) Cisplatin monoaqua-substituted derivative; (iii)

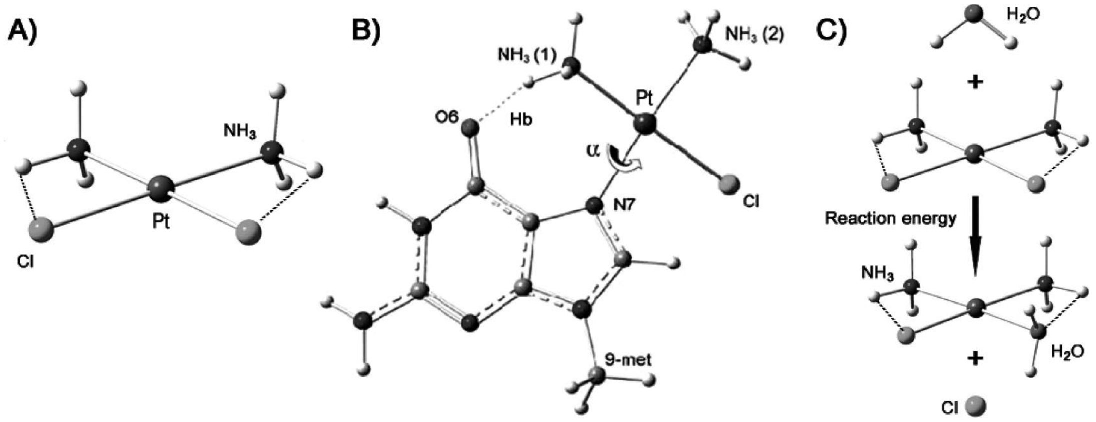


Figure 1. Schematic view of the three platinated systems studied: (A) Cisplatin; (B) Cisplatin−9metG adduct; (C) reactants and products of the first aquation process of Cisplatin. Intramolecular and intermolecular hydrogen bonds are shown with dashed lines.

Cisplatin 9-methyl-guanine adduct; and (iv) the first aquation process of Cisplatin.

Theoretical Methods

The structures of Cisplatin, 9-methyl-guanine (9metG), water, and the monoaquo complex shown in Figure 1 were fully optimized as isolated species minima, at different levels of theory using gradient techniques, without imposing any structural or symmetry constraints.

Calculations using Gaussian basis sets were performed at the HF, MP2[10] (full and frozen core), and DFT levels using the GGA functional PBE1PBE[11a] and one- to three-parameter hybrid functionals (mPW1PW91,[11b] B3LYP,[11c,d] B3PW91,[11c,e] and B3P86[11c,f]) as implemented in Gaussian03, revision B05.[9] Two ECP schemes, constituted by a pseudopotential and the concomitant double-ζ (Dζ) quality basis set, were applied for Pt as follows: (1) Stevens, Basch, and Krauss (SBK)[12a,b] ECP with CEP-31G that leads to the (7s7p5d)/[4s4p3d]-GTO valence basis set for Pt and (2) Los Alamos National Laboratory's LanL2DZ ECP developed by Hay and Wadt,[13a–d] that employs the (8s6p3d)/[5s3p2d]-GTO valence basis set for Pt. The same Pople-type basis set with polarization functions 6-31G(d)[14] is used for ligands in Cisplatin (NH_3 and Cl^-) and derivatives (H_2O and 9-methyl-guanine). The nature of the stationary points—minima—was verified at each level of theory considered here, based on the analysis of the corresponding analytical Hessian matrix.

QM computations using numerical basis sets were performed at the DFT level with SIESTA code.[7] SIESTA has shown an excellent performance for medium and large systems and has also proven to be appropriate for biomolecules and metal ions in biological systems.[15] SIESTA reads the norm-conserving pseudopotentials in semilocal form (a different radial potential $V(r)$ for each angular momentum l), generally using the Troullier−Martins parametrization.[16] Then, it transforms this semilocal form into the fully nonlocal form proposed by Kleinman and Bylander (KB).[7] The use of standard norm-conserving pseudopotentials[16] avoids the computation of core electrons, smoothing at the same time the valence charge density. Pt is treated as an 18-electron system, namely with both $n = 5$ and $n = 6$ considered as valence electron shells (Table 1 reports the reference valence configurations, cutoff radius, and total number of KB projectors used for each atom). SIESTA uses basis set functions that consist of localized (numerical) pseudoatomic orbitals (PAO), which are projected on a real space grid to compute the Hartree potential and exchange-correlation potential's matrix elements. Dζ plus polarization quality basis sets were employed for all atoms, with a PAO energy shift of 20 meV and a grid cutoff of 200 Ry.[8] To improve the thermodynamic characterization of species participating in Cisplatin aquation process, single-point calculations over the structure of the isolated species minima were carried out with a PAO energy shift of 0.5 meV and a grid cutoff of 300 Ry. Calculations were performed using the Perdew, Burke, and Ernzerhof GGA functional[10] (PBE_{SIESTA}), which coincides with the PBE1PBE functional implemented in Gaussian03 (G03).

Table 1. Reference Valence Configurations, Cutoff Radius, and Total Number of KB Projectors Used in the Troullier−Martins Procedure to Obtain the Pseudopotentials for Each Atom

atom	valence configuration	cutoff radius	no. of KB projectors
H	1s(1.00)	1.25	9
	2p(0.00)		
	3d(0.00)		
C	2s(2.00)	1.25	16
	2p(2.00)		
	3d(0.00)		
	4f(0.00)		
N	2s(2.00)	1.20	16
	2p(3.00)		
	3d(0.00)		
	4f(0.00)		
O	2s(2.00)	1.13	16
	2p(4.00)		
	3d(0.00)		
	4f(0.00)		
Cl	3s(2.00)	1.50	16
	3p(5.00)		
	3d(0.00)		
	4f(0.00)		
Pt	6s(1.00)	2.47	16
	6p(0.00)	2.87	
	5d(9.00)	1.98	
	5f(0.00)	2.30	

The energy of reaction has been calculated as the subtraction between the sum of the products and the sum of the

Table 2. Calculated Cisplatin Bond Lengths (Å) and Angles (deg) at Several Levels of Theory, OMPBDs, and OMPGDs with Respect to Experimental Values and Best Available Theoretical Data

	Pt−Cl	Pt−N	N−Pt−N	N−Pt−Cl	Cl−Pt−Cl	OMPBD	OMPGD
			Pseudopotential and Basis Set Used for Pt: LANL2/LANL2DZ				
PBE1PBE	2.322 (−0.008)	2.086 (0.076)	98.2 (11.2)	83.3 (−7.0)	95.1 (3.2)[a]	2.1 [2.2][e]	5.6 [4.2]
mPW1PW91	2.324 (−0.006)	2.089 (0.079)	98.1 (11.1)	83.4 (−6.9)	95.1 (3.2)	2.1 [2.2]	5.6 [4.2]
B3LYP	2.349 (0.019)	2.110 (0.100)	98.2 (11.2)	83.3 (−7.0)	95.3 (3.4)	2.9 [2.2]	6.0 [4.3]
B3PW91	2.330 (0.000)	2.090 (0.080)	98.3 (11.3)	83.3 (−7.0)	95.1 (3.2)	2.0 [2.1]	5.6 [4.2]
B3P86	2.325 (−0.005)	2.090 (0.080)	98.4 (11.4)	83.2 (−7.1)	95.2 (3.3)	2.1 [2.2]	5.8 [4.3]
mp2(full)	2.345 (0.015)	2.093 (0.083)	97.2 (10.2)	83.8 (−6.5)	95.1 (3.2)	2.4 [1.9]	5.4 [3.8]
MP2(FC)[b]	2.347 (0.017)	2.083 (0.073)	97.4 (10.4)	83.9 (−6.4)	94.8 (2.9)	2.2 [1.6]	5.3 [3.6]
MP3(FC)[b]	2.358 (0.028)	2.098 (0.088)	96.9 (9.9)	83.9 (−6.4)	95.3 (3.4)	2.8 [1.7]	5.5 [3.7]
MP4(FC)[b]	2.358 (0.028)	2.095 (0.085)	97.1 (10.1)	84.0 (−6.3)	95.0 (3.1)	2.7 [1.6]	5.5 [3.6]
HF	2.363 (0.033)	2.124 (0.114)	95.3 (8.3)	84.5 (−5.8)	95.7 (3.8)	3.5 [2.4]	5.4 [3.5]
			Pseudopotential and Basis Set Used for Pt: SBK/CEP-31G				
PBE1PBE	2.308 (−0.022)	2.093 (0.083)	98.0 (11.0)	83.5 (−6.8)	95.0 (3.1)	2.5 [2.7]	5.7 [4.3]
mPW1PW91	2.310 (−0.020)	2.095 (0.085)	98.0 (11.0)	83.5 (−6.8)	95.0 (3.1)	2.5 [2.7]	5.7 [4.3]
B3LYP	2.337 (0.007)	2.123 (0.113)	98.1 (11.1)	83.4 (−6.9)	95.2 (3.3)	3.0 [2.8]	6.0 [4.5]
B3PW91	2.316 (−0.014)	2.101 (0.091)	98.1 (11.1)	83.4 (−6.9)	95.0 (3.1)	2.6 [2.7]	5.8 [4.4]
B3P86	2.312 (−0.018)	2.095 (0.085)	98.3 (11.3)	83.3 (−7.0)	95.1 (3.2)	2.5 [2.6]	5.8 [4.4]
mp2(full)	2.307 (−0.023)	2.098 (0.088)	96.8 (9.8)	84.2 (−6.1)	94.7 (2.8)	2.7 [2.8]	5.3 [3.9]
MP2(FC)[b]	2.312 (−0.018)	2.090 (0.080)	97.1 (10.1)	84.2 (−6.1)	94.6 (2.7)	2.4 [2.5]	5.2 [3.8]
MP3(FC)[b]	2.327 (−0.003)	2.106 (0.096)	96.2 (9.2)	84.3 (−6.0)	95.2 (3.3)	2.5 [2.6]	5.1 [3.7]
MP4(FC)[b]	2.326 (−0.004)	2.105 (0.095)	96.7 (9.7)	84.2 (−6.1)	94.8 (2.9)	2.4 [2.6]	5.2 [3.8]
HF[b]	2.348 (0.018)	2.140 (0.130)	95.0 (8.0)	84.7 (−5.6)	95.7 (3.8)	3.6 [3.0]	5.4 [3.6]
			Pseudopotential and Basis Set Used for Pt: Troullier−Martins/Dζ				
PBE_{SIESTA}	2.331 (0.001)	2.096 (0.086)	100.3 (13.3)	81.6 (−8.7)	96.5 (4.6)	2.2 [2.2]	6.8 [5.4]
			Experimental Values (Mean and Range)[c]				
exp	2.330	2.010	87.0	90.3	91.9	0.0	0.0
exp	2.328–2.333	1.950–2.050	85.5–88.5	88.5–92.0	91.3–92.2		
			Car–Parrinello Molecular Dynamics at Room Temperature[d]				
BLYP	2.36	2.03	89	88	94	[0.0]	[0.0]

[a] Values in parentheses correspond to the difference between the calculated and the mean experimental value (Δ_{c-o}) for each structural parameter. [b] Taken from ref 3d. [c] Taken from ref 17. [d] Calculated average values taken from ref 4d. [e] Values in square brackets correspond to OMPBD and OMPGD calculated with respect to the CPMD-BLYP simulation.

isolated reactants. The convenience of reporting the reaction energy considering the first aquation process as a bimolecular reaction is discussed elsewhere.[5m] In our work, BSSE was not considered.

Results and Discussion

Tables 2−6 collect representative structural parameters for Cisplatin, its monoaqua-substituted derivative, and the Cisplatin−9metG adduct as determined with each theoretical level using different combinations of functionals and ECP/basis sets (LANL2DZ, SBK and Troullier−Martins schemes). Since the aim of this work is comparing DFT implementations using analytical and numerical basis sets as respectively present in G03 and SIESTA, taking as reference the available experimental structural data,[17–19] differences between predicted and observed values (labeled as Δ_{c-o}) have been calculated and reported for all bonds and angles involving heavy-atoms. In addition, the overall mean percentage of all bond differences (OMPBD) and the overall mean percentage of all structural parameters differences (OMPGD) were calculated using Δ_{c-o} as follows:

$$\mathrm{OMPBD} = \left[\sum_{i=\mathrm{bond}}^{n} \frac{|\Delta_{\mathrm{c-o}}|_i \times 100}{(\mathrm{BD}_{\mathrm{obsvd}})_i}\right]\frac{1}{n} \qquad (1)$$

$$\mathrm{OMPGD} = \left[\sum_{i=\mathrm{bond}}^{n} \frac{|\Delta_{\mathrm{c-o}}|_i \times 100}{(\mathrm{BD}_{\mathrm{obsvd}})_i} + \sum_{j=\mathrm{angle}}^{m} \frac{|\Delta_{\mathrm{c-o}}|_j \times 100}{(\mathrm{BA}_{\mathrm{obsvd}})_j}\right]\left(\frac{1}{n+m}\right) \qquad (2)$$

BD_{obsvd} and BA_{obsvd} being respectively the observed experimental values of each of the *n* bond distances taken into account and each of the *m* bond angles considered in calculating OMPGD.

To test the performance of numerical basis set and the Troullier−Martins pseudopotential scheme against energetic parameters, reaction energy for the first aquation process of Cisplatin was calculated from the isolated species energetics and reported in Table 7. Since there is no experimental data for this reaction in gas phase, our calculations are compared to other theoretical determinations.[5b,c]

Geometry of Cisplatin. The first aspect to be taken into account is that experimental data of reference for Cisplatin come from X-ray diffraction studies in the solid state obtained by Milburn et al.[17] Due to packing interactions, distortions on the structural parameter from the gas phase are likely to be present in the solid state, hence no perfect agreement with X-ray crystallographic results are expected, even for the best theoretical method employed to determine the structure of isolated Cisplatin. Actually, the intermolecular interaction between two adjacent Cisplatin molecules present in the crystal indicated the formation of two hydrogen bonds from each nitrogen of one Cisplatin molecule, both bonds being to the same chloride atom in the next molecule along the *c*-axis of the unit cell (donor–acceptor H bond distance of 3.3 Å).[17] This H bond interaction leaves both hydrogen atoms lying in the plane of the molecule in the N−Pt−N quadrant and produces a loss of symmetry given a quasi C_{2v} conformation different than that calculated as an isolated species minimum. As a consequence, the two

Table 3. Calculated Cisplatin Monoaquo Bond Lengths (Å) and Angles (deg) at Several Levels of Theory, OMPBDs, and OMPGDs with Respect to Best Available Theoretical Data

	Pt−Cl	Pt−N[a]	Pt−O	N−Pt−N	N−Pt−Cl	N−Pt−O	O−Pt−Cl	OMPBD	OMPGD
				Pseudopotential and Basis Set Used for Pt: LANL2/LANL2DZ					
PBE1PBE	2.30 (−0.03)	2.03 (0.00)	2.10 (0.00)	97 (6)	87 (−1)	87 (−7)	89 (3)[b]	0.4	2.9
mPW1PW91	2.30 (−0.03)	2.03 (0.00)	2.11 (0.01)	97 (6)	87 (−1)	87 (−7)	89 (3)	0.6	2.9
B3LYP	2.33 (0.00)	2.05 (0.02)	2.13 (0.03)	97 (6)	87 (−1)	87 (−7)	89 (3)	0.8	3.0
B3PW91	2.31 (−0.02)	2.03 (0.00)	2.11 (0.01)	97 (6)	88 (0)	87 (−7)	89 (3)	0.4	2.7
B3P86	2.30 (−0.03)	2.03 (0.00)	2.10 (0.00)	97 (6)	87 (−1)	87 (−7)	89 (3)	0.4	2.9
mp2(full)	2.32 (−0.01)	2.04 (0.01)	2.11 (0.01)	96 (5)	88 (0)	87 (−7)	89 (3)	0.5	2.5
MP2(FC)	2.32 (−0.01)	2.04 (0.01)	2.11 (0.01)	96 (5)	88 (0)	87 (−7)	89 (3)	0.5	2.5
HF	2.33 (0.00)	2.07 (0.04)	2.12 (0.02)	95 (4)	88 (0)	88 (−6)	89 (3)	1.0	2.5
				Pseudopotential and Basis Set Used for Pt: SBK/CEP-31G					
PBE1PBE	2.28 (−0.05)	2.03 (0.00)	2.10 (0.00)	96 (5)	87 (−1)	87 (−7)	89 (3)	0.7	2.8
mPW1PW91	2.29 (−0.04)	2.04 (0.01)	2.10 (0.00)	96 (5)	87 (−1)	87 (−7)	89 (3)	0.7	2.8
B3LYP	2.31 (−0.02)	2.06 (0.03)	2.13 (0.03)	96 (5)	87 (−1)	87 (−7)	89 (3)	1.3	3.0
B3PW91	2.29 (−0.04)	2.04 (0.01)	2.11 (0.01)	96 (5)	87 (−1)	87 (−7)	89 (3)	0.9	2.9
B3P86	2.29 (−0.04)	2.04 (0.01)	2.10 (0.00)	97 (6)	87 (−1)	87 (−7)	89 (3)	0.7	3.0
mp2(full)	2.29 (−0.04)	2.04 (0.01)	2.10 (0.00)	96 (5)	88 (0)	87 (−7)	90 (4)	0.7	2.8
MP2(FC)	2.29 (−0.04)	2.04 (0.01)	2.10 (0.00)	96 (5)	88 (0)	87 (−7)	90 (4)	0.7	2.8
HF	2.32 (−0.01)	2.08 (0.05)	2.13 (0.03)	95 (4)	89 (1)	88 (−6)	89 (3)	1.4	2.8
				Pseudopotential and Basis Set Used for Pt: Troullier−Martins/Dζ					
PBE$_{\text{SIESTA}}$	2.31 (−0.02)	2.03 (0.00)	2.12 (0.02)	97 (6)	86 (−2)	88 (−6)	89 (3)	0.6	2.9
				Car–Parrinello Molecular Dynamics at Room Temperature[c]					
BLYP	2.33	2.03	2.10	91	88	94	86	0.0	0.0

[a] Amino group trans to water. [b] Values in parentheses correspond to the difference calculated respect to the corresponding structural parameter extracted form the best theoretical calculation known (Δ_{c-o}). [c] Calculated average values taken from ref 4d.

N−H•••Cl intramolecular hydrogen bonds that appear in the gas phase (see dashed lines in Figure 1A) are substituted by two intermolecular hydrogen bonds in the crystal structure forcing the closure of the N−Pt−N angle.

For completeness and uniformity with the structural analysis of the monoaquo complex (see Table 3 and the corresponding discussion), our results for Cisplatin are also compared with the DFT Car–Parrinello molecular dynamics (CPMD-BLYP) simulation at room temperature, using periodic boundary conditions and 35 explicit water molecules to include solvent effects.[4d] The corresponding OMPBDs and OMPGDs values are reported in square brackets in Table 2.

Comparative analysis of the calculated Pt−Cl bond lengths in Cisplatin collected in Table 2 shows that whereas LanL2DZ ECP calculations essentially tend to overestimate this parameter with respect to the corresponding mean experimental values—coming from different Cisplatin molecules present in the triclinic X-ray unit cell[17]—using SBK means to underestimate it. Concerning DFT functionals and basis sets, the best results are obtained with PBE as implemented in SIESTA, as well as with the hybrid functional B3PW91 using LanL2DZ ECP for Pt. Regarding the calculated Pt−N bonds, all the methods exhibit the same trend, overestimating these distances, being the best results obtained at the MP2(FC)/LanL2DZ level, followed by both PBE1PBE/LanL2DZ and mPW1PW91/LanL2DZ, which are slightly more accurate than PBE$_{\text{SIESTA}}$, but still showing a very good performance comparable to that achieved at the MP4(FC) level using LanL2DZ ECP. Thus, in general terms, the LanL2DZ pseudopotential produces better results than SBK in predicting bond lengths, as reflected by the corresponding OMPBD values collected in Table 1 (B3PW91 = 2.0, PBE1PBE = 2.1, mPW1PW91 = 2.1, B3P86 = 2.1, and MP2(FC) = 2.2). In comparing results obtained with LanL2DZ and GTOs in G03 with those obtained using a numerical basis set in SIESTA, it is shown that PBE$_{\text{SIESTA}}$ achieves a performance similar to MP2(FC) and with the DFT calculations already mentioned, but with a significantly lower computational effort. It is worthy to notice that the largest OMPBDs are obtained at the HF level indicating the importance of the inclusion of a dynamic electron correlation in the calculation of bond lengths, a fact previously reported for the Pt−Cl bond.[3d]

Considering now N−Pt−N and Cl−Pt−Cl bond angles, it can be seen that compared with experiment they are systematically overestimated—leading thus to underestimation of the N−Pt−Cl angle—by all methods disregarding any variation on the corresponding ECPs and basis set. None of the calculated bond angles fell within the experimental range. OMPGD and OMPBD values reported in Table 2 clearly show that the general performance is lower in quality when modeling angles respect to prediction of bond distances, a fact already noticed in every methodological comparison previously reported.[3d,e,5e] The Cl−Pt−Cl angle is somehow better modeled than N−Pt−N, with deviations from experiment ranging from 2.7° to 4.6°. SBK ECP gives a global better performance than LanL2DZ. In all the cases, the lack of general agreement between the observed angles and the calculated ones in gas phase are due to the already mentioned effects of packing a problem that is attenuated when comparing the data to the best available calculated structures in aqueous solution. Thus, OMPGD is only a qualitative tool at the moment of determining the global structure obtained with the different methods showing in all the cases a difference ranging from 5 to 7% (or 3.5–5.4%, depending on the nature of the data considered as a reference).

OMPBDs and OMPGDs calculated with respect to CPMD-BLYP data are close (and slightly smaller, in general terms, see Table 2) to those obtained taking X-ray crystallographic results as reference, reflecting the proximity of the structural parameters obtained by optimization of the isolated species at 0° K to the data coming from the average over aqueous solution dynamics performed at room temperature. The principal trends already discussed are sustained in considering these values. Nevertheless, two points emerge from the analysis: (a) Möller–Plesset levels of theory become the best in predicting bond lengths for solution structures; (b) the best OMPGDs are now obtained at the HF level, probably by compensation of errors leading to a better characterization of the angles between NH_3 groups.

Geometry of Cisplatin Monoaqua-Substituted Derivative. As far as we know, there is no experimental characterization of the molecular structure of the monoaquo complex derivated from Cisplatin. The cationic complex $[Pt(H_2O)_4]^{2+}$ is the most related chemical species for which structural data obtained from EXAFS are available.[5a] The mean Pt−O bond distance reported for the tetra-aquo compound in C_{2v} conformation is 2.01 Å,[19] a value that can be thought as a semiquantitative reference for the Pt−O bond in the monoaquo species. To achieve the comparison of all bond lengths and angles that involve the heavy atoms in the monoaquo complex we have taken as a reference the values from a CPMD-BLYP simulation at room temperature, using periodic boundary conditions and 35 explicit water molecules, as previously done for Cisplatin.[4d] So we are now comparing the structural parameters calculated for isolated species with the simulated ones in aqueous solution, and thus, OMPBDs and OMPGDs collected in Table 3 were calculated using the results coming from the DFT Car–Parrinello molecular dynamics as the unique reference.

All three calculated bond lengths scarcely deviate from the reference values, showing for all methods a deviation that lies in the hundredths of angstroms. Results obtained with LanL2DZ are slightly better that the corresponding SBK ECP ones. For bond length parameters, PBE_{SIESTA} still shows a very good performance, comparable to that of mPW1PW91/LanL2DZ, with an overall agreement to the reference data (OMPBD = 0.6) placed in between the lowest quality set of results (obtained using analytical GTOs and the ECP of SBK, with OMPBDs in the 0.7–1.4 range) and the best ones achieved at the DFT/LanL2DZ levels (the lowest OMPBD value of 0.4 was obtained using PBE1PBE, B3PW91, and B3P86 functionals). Again, the small ranges spanned by all methodologies indicate that bond lengths from isolated species are fairly close to those from aqueous solution simulations.

In all cases, there is a better agreement in the two calculated bond angles involving a Cl^- ion (N−Pt−Cl and O−Pt−Cl). All methods overestimate the N−Pt−N angle, whereas the N−Pt−O angle is always underestimated a few degrees. OMPGD values ranged for all DFT functionals from 2.7 to 3.0 (being the best of them B3PW91/LanL2DZ to be compared to the best performance of 2.5 obtained at the MP2/LanL2DZ level and to a PBE_{SIESTA} performance of 2.9) showing again the excellent agreement between isolated species optimization and solvated CPMD-BLYP.

Geometry of Cisplatin−9metG Adducts. Also in this case, experimental values available for comparison come from X-ray data. As shown in Figure 2, our structural data obtained by optimization of isolated species under gas phase conditions are compared to data from three monofunctional adducts formed between Cisplatin (or some aquo complex derivative) and a double stranded B-DNA dodecamer of sequence d{5′-CpGpCpG*pApApTpTpCpG*pCpG-3′}{5′-CpG*pCpGpApApTpTpCpGpCpG-3′}, where the asterisk marks the sites of covalent binding (first step of platination).[18] The first inspection of the crystal structure shows the formation of three adducts in different neighbor contexts of nucleobases. Two bonds are of the type 5′-CG*C-3′, and one is of the type 5′-CG*A-3′. The binding of Cisplatin to guanine residues in different DNA sequences clearly affects, in terms of structural parameters, the formation of the Cisplatin−guanine adducts.[7] As a counterpart, the structural damage produced by platination to DNA also depends upon the specific nucleobase sequence.[7] Consistently, these aspects must be kept in mind when analyzing the experimental range and mean values reported in Table 6.

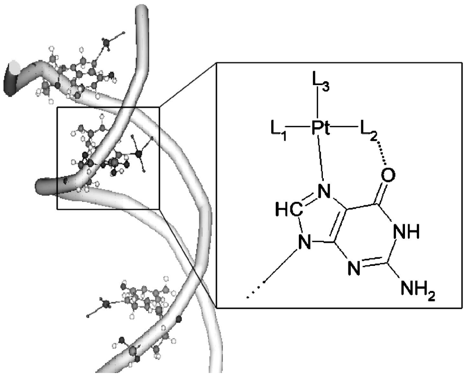


Figure 2. Schematic representation of the crystal structure (resolution: 2.6 Å) of the primary mode of binding of Cisplatin to the B-DNA dodecamer by Wing[18] et al. The PDB file (DDL017) was obtained from the Nucleic Acid Database (http://ndbserver.rutgers.edu/). The amplified sketch shows one of the formed adducts in which Cisplatin ligands are labeled as in ref 18.

A second inspection of the X-ray structure demonstrates that the authors were not capable of determining the chemical nature of the ligands of the platinum complex (labeled L1−L3 in Figure 2).[18] It has been hypothesized that the L2 ligand could be a water molecule suggesting that the diaqua-substituted derivative was the platination agent, before crystallization was achieved.[18] Nevertheless, the knowledge of the conditions of crystallization (solution of 2-methyl-2,4-pentanediol a 4 °C)[18] and of further studies leading to determine the equilibrium relation in aqueous solution of Cisplatin, monoaquo, monohydroxo, diaquo, and dihydroxo complexes at 37 °C and different concentrations of chloride[20] allowed us to reassign the three unknown ligands as follow: L1 = Cl^- (labeled as X in Tables 4−6); L2 = N(1); and L3 = N(2).

Pt−NH_3 bond lengths calculated with all methods are in very good agreement with the experimental values taken as reference. While Pt−N(2) is always a little bit underestimated, Pt−N(1) is always overestimated by a few hundredths of angstroms. Pt−N7 bond lengths are not so well-modeled, and none of the calculated values fell within the experimental range, being for all methods shorter than expected. Calculated

Table 4. *cis*-[Pt(NH_3)$_2$Cl(9-met-guanine)]$^+$ Structure—Distances in angstroms; Angles in degress—Calculated at Several Levels of Theory Using LanL2/LanL2DZ[a]

	LANL2/LANL2DZ							
	PBE1PBE	mPW1PW91	B3LYP	B3PW91	B3P86	MP2(FC)	MP2(Full)	HF
Pt-X[b]	2.329	2.331	2.356	2.336	2.332	2.346	2.345	2.367
Pt−N(2)	2.066 (−0.055)[c]	2.069 (−0.052)	2.092 (−0.029)	2.074 (−0.047)	2.067 (−0.054)	2.085 (−0.036)	2.085 (−0.036)	2.106 (−0.015)
Pt−N(1)	2.078 (0.023)	2.081 (0.026)	2.109 (0.054)	2.088 (0.033)	2.082 (0.027)	2.082 (0.027)	2.082 (0.027)	2.130 (0.075)
Pt−N7	2.031 (−0.206)	2.033 (−0.204)	2.057 (−0.180)	2.037 (−0.200)	2.032 (−0.205)	2.024 (−0.213)	2.023 (−0.214)	2.081 (−0.156)
N(1)···O (Hb)	1.745	1.750	1.766	1.753	1.731	1.833	1.831	1.892
X[b]−Pt−N(2)	85.2 (−14.4)	85.2 (−14.4)	85.0 (−14.6)	85.2 (−14.4)	85.1 (−14.5)	85.9 (−13.7)	85.9 (−13.7)	86.2 (−13.4)
X[b]−Pt−N7	90.2 (−5.9)	90.2 (−5.9)	90.3 (−5.8)	90.2 (−5.9)	90.2 (−5.9)	89.7 (−6.4)	89.7 (−6.4)	90.3 (−5.8)
N(2)−Pt−N(1)	93.6 (11.9)	93.6 (11.9)	93.5 (11.8)	93.7 (12.0)	93.7 (12.0)	93.8 (12.1)	93.8 (12.1)	94.5 (12.8)
N(1)−Pt−N7	91.0 (10.4)	90.9 (10.3)	91.1 (10.5)	90.9 (10.3)	91.0 (10.4)	90.6 (10.0)	90.6 (10.0)	91.0 (10.4)
N(1)···O (Hb)	163	163	163	163	164	158	158	156
α	37 (11)	37 (11)	36 (10)	37 (11)	36 (10)	41 (15)	41 (15)	37 (11)
OMPBD	4.3	4.3	4.0	4.3	4.3	4.2	4.2	3.8
OMPGD	12.9	12.9	12.3	12.9	12.5	14.7	14.7	12.7

[a] Performance with respect to experimental data reported as deviations, OMPBDs and OMPGDs. [b] The nature of this ligand is not defined in the work by Wing et al.,[18] but it can be thought of as a chloride. [c] For comparison, we show in parentheses the difference between the calculated value and the mean experimental or observed value (Δ_{c-o}) taken from ref 18. See Table 6 for experimental values.

Table 5. *cis*-[Pt(NH_3)$_2$Cl(9-met-guanine)]$^+$ Structure—Distances in angstroms; Angles in deg—Calculated at Several Levels of Theory Using SBK/CEP-31G ECP[a]

	SBK/CEP-31G							
	PBE1PBE	mPW1PW91	B3LYP	B3PW91	B3P86	MP2(FC)	MP2(Full)	HF
Pt−X[b]	2.311	2.313	2.340	2.319	2.315	2.306	2.305	2.348
Pt−N(2)	2.067 (−0.054)[c]	2.070 (−0.051)	2.092 (−0.029)	2.075 (−0.046)	2.070 (−0.051)	2.074 (−0.047)	2.074 (−0.047)	2.110 (−0.011)
Pt−N(1)	2.088 (0.033)	2.090 (0.035)	2.116 (0.061)	2.096 (0.020)	2.090 (0.015)	2.089 (0.034)	2.089 (0.034)	2.128 (0.073)
Pt−N7	2.041 (−0.196)	2.043 (−0.194)	2.068 (−0.169)	2.047 (−0.141)	2.042 (−0.147)	2.029 (−0.208)	2.028 (−0.209)	2.093 (−0.144)
N(1)···O (Hb)	1.749	1.753	1.769	1.760	1.734	1.836	1.834	1.898
X[b]−Pt−N(2)	85.4 (−14.2)	85.4 (−14.2)	85.2 (−14.4)	85.3 (−14.3)	85.3 (−14.3)	86.4 (−13.2)	86.4 (−13.2)	86.3 (−13.3)
X[b]−Pt−N7	90.3 (−5.8)	90.3 (−5.8)	90.3 (−5.8)	90.3 (−5.8)	90.2 (−5.9)	89.8 (−6.3)	89.7 (−6.4)	90.3 (−5.8)
N(2)−Pt−N(1)	93.6 (11.9)	93.6 (11.9)	93.5 (11.8)	93.7 (12.0)	93.6 (11.9)	93.7 (12.0)	93.7 (12.0)	92.5 (10.8)
N(1)−Pt−N7	90.8 (10.2)	90.8 (10.2)	91.0 (10.4)	90.7 (10.1)	90.9 (10.3)	90.1 (9.5)	90.2 (9.6)	90.9 (10.3)
N(1)···O (Hb)	163	163	164	164	164	158	158	156
α	37 (11)	37 (11)	36 (10)	37 (11)	36 (10)	42 (16)	41 (15)	37 (11)
OMPBD	4.3	4.3	4.0	4.2	4.3	4.4	4.4	3.5
OMPGD	12.8	12.8	12.3	12.8	12.4	15.1	14.7	12.3

[a] Performance with respect to experimental data reported as deviations, OMPBDs and OMPGDs. [b] The nature of this ligand is not defined in the work by Wing et al.,[18] but it can be thought of as a chloride. [c] For comparison, we show in parentheses the difference between the calculated value and the mean experimental or observed value (Δ_{c-o}) taken from ref 18. See Table 6 for experimental values.

Table 6. *cis*-[Pt$(NH_3)_2$Cl(9-met-guanine)]$^+$ Structure—Distances in angstroms; Angles in deg—Calculated with SIESTA at the PBE$_{SIESTA}$ Level of Theory with Troullier−Martins Pseudopotentials, Deviations, OMPBDs and OMPGDs, and Reference Experimental Data

	Troullier-Martins/Dζ		
	PBE$_{SIESTA}$	exp (mean)	exp (range)
Pt−X[a]	2.329		
Pt−N(2)	2.066(−0.055)[b]	2.121	1.999–2.230
Pt−N(1)	2.078(0.023)	2.055	1.814–2.247
Pt−N7	2.031(−0.206)	2.237	2.164–2.315
N(1)···O(Hb)	1.745		
X[a]−Pt−N(2)	85.2(−14.4)	99.6	95.5–104.7
X[a]−Pt−N7	90.2(−5.9)	96.1	79.0–105.6
N(2)−Pt−N(1)	93.6(11.9)	81.7	78.7–85.1
N(1)−Pt−N7	91.0(10.4)	80.6	69.7–102.0
N(1)···O(Hb)	163		
α	37 (11)	26	13–44
OMPBD	4.3		
OMPGD	12.9		

[a] The nature of this ligand is not clearly defined in the work by Wing et al.,[18] but it can be thought of as a chloride. [b] For comparison purposes, difference between the calculated value and the corresponding mean experimental or observed value (Δ_{c-o}) taken from ref 18 is reported in parenthesis.

Table 7. Calculated Reaction Energy (kcal/mol) for the First Aquation Process of Cisplatin at the Different Levels of Theory over Optimized Isolated Species in the Gas Phase

	ECP/basis set employed		
level of theory	LANL2/ LANL2DZ	SBK/ CEP-31G	Troullier− Martins/Dζ
PBE1PBE/ PBE$_{SIESTA}$	115 (−5)	116 (−4)	142 (22)[a] 127 (7)[b]
mPW1PW91	115 (−5)	116 (−4)	
B3LYP	114 (−6)	115 (−5)	
B3PW91	115 (−5)	116 (−4)	
B3P86	119[c] (−1)	116 (−4)	
MP2(FC)	116[c] (−4)	116 (−4)	
mp2(full)	113 (−7)	115 (−5)	
HF	108 (−12)	109 (−11)	
G3-type strategy[d]	120		

[a] Results obtained from optimized structures using a PAO energy shift of 20 meV and a grid cutoff of 200 Ry. [b] Results obtained from single-point calculations over optimized structures using a PAO energy shift of 0.5 meV and a grid cutoff of 300 Ry. [c] From ref 5b. [d] For comparison purposes, differences between our calculated values and the best known theoretical value (G3-type strategy) taken from ref 5c are reported in parenthesis.

OMPBDs ranging from 3.5 to 4.4 showed good performance of all methods, HF (3.5/3.8) and B3LYP (4.0) being the most accurate ones, regardless the ECP used. PBE$_{SIESTA}$ (OMPBD = 4.3) also gives a good overall result, comparable to the one obtained with PBE1PBE and the other hybrid functionals using analytical basis sets. MP2 (full and FC) with the ECP of SBK gave the poorest results, particularly in the description of the Pt−N7 bond.

A measure of the strength of the hydrogen bond (Hb) formed between N(1) and the O6 carbonyl group of guanine could not be obtained from the X-ray experiment since hydrogen atoms were not detected. The formation of this Hb is thought to be essential for the stabilization of the first adduct formed between Cisplatin and DNA.[6q] A recent work by van der Wijst et al., in which the performance of various density functionals in describing the hydrogen bonds in DNA base pairs was analyzed, shows that GGA functionals BP86 and PW91 gives the best results compared to X-ray values, while B3LYP tends to underestimate the hydrogen bond strength compared to experiment.[21] In our work, B3P86 gave the strongest Hb (shorter bond distance and better alignment among the three atoms involved), and if this is taken as a reference, PBE$_{SIESTA}$ and PBE1PBE/LanL2DZ gave the following better results.

For the calculated angles, all methods give the same trends: the X−Pt−N(2) angle is systematically underestimated while the N(2)−Pt−N(1) angle is always overestimated by more than 10° and never fell within the experimental range. The X−Pt−N7 and N(1)−Pt−N7 angles are better modeled, particularly the first one, giving additional support to the reassignment of L2 as a chloride. The lowest OMPGDs in a range of 12.3–15.1 are obtained by B3LYP and B3P86 regardless of the ECP used (values of 12.3 and 12.4/12.5, respectively). For the structural characterization of the Cisplatin−9metG adduct, LanL2DZ and SBK offer the same level of accuracy. PBE$_{SIESTA}$, again, gives a very good result, with an accuracy comparable to PBE1PBE and the others functionals (OMPGD = 12.9). If global differences are analyzed, MP2 shows the largest deviation with respect to the experimental information. This is an outcome of a worse representation of the dihedral angle labeled α (Figure 1B), which is greatly determined by the Hb formed with the carbonyl group of guanine. A good modeling of this dihedral, that represents the angle formed between the plane of the platinum complex and the plane of the nucleobase, is crucial to estimate the stability of the adduct and the relative orientation toward the formation of the bifunctional complex (all DFT functionals and HF gave a deviation of ∼10 Å from the experimental mean value).

Energetics of the First Aquation Process of Cisplatin. A good description of the reaction energy and activation energy is decisive to understand the thermodynamic and kinetic behaviors of the reactions in which Cisplatin and its derivatives are involved. As far as we know, the best thermodynamic result obtained for the first aquation process of Cisplatin was achieved using a G3 modified strategy.[5c] The G3 modified treatment (which includes a term for describing spin–orbit coupling, uses the MWB-60 pseudopotential for the Pt and CCSD(T) instead of the original QCISD(T), replacing the G3Large basis set with the more common aug-cc-pvtz)[5c] was used as a reference to check the performance of our theoretical results when predicting the reaction energy.

As we can see in Table 7, B3P86/LanL2DZ reported by Chval and Sip[5b] gives an extraordinary good result if compared to G3. If the general accuracy of the methods used in the energetic characterization is kept in the aquation reaction of Cisplatin, we can presume that the G3 calculation does not deviate more than 1.0 kcal/mol and that the GGA and hybrid functionals accuracy rounds approximately between 3.5 and 7.5 kcal/mol apart from the probable experimental result. This is exactly the behavior that is reproduced by the selected DFT methods that use analytical basis set functions in comparison with G3. In contrast, the results obtained from optimized species with PBE_{SIESTA} using a PAO energy shift of 20 meV and a grid cutoff of 200 Ry overestimate the calculated reaction energy by at least 22 kcal/mol leading to a worse result than HF. Nevertheless, the qualitative reaction profile is still the same in all the cases, being the first aquation process of Cisplatin in the gas phase largely endothermic.

However, if the quality of the SIESTA calculation is improved, the PBE_{SIESTA} result deviates in absolute value only 7 kcal/mol from the G3 strategy, a similar performance that MP2(full) calculations with the analytical basis functions. A comparison of PBE_{SIESTA} with PBE1PBE as implemented in G03, lead us to deduce that the Troullier−Martins pseudopotential and the numerical basis set of SIESTA also give quantitatively good results in predicting the reaction energy.

Conclusions

Our results show that SIESTA gives geometrical parameters of very good to excellent accuracy for the complexes of platinum considered herein. Particularly good results are obtained for the geometry of the Cisplatin−9metG adducts, allowing us to extend the calculations (having the performance of SIESTA in mind) to larger and more complex molecular systems. Energetic results of SIESTA show a qualitative good agreement with more standard implementations of DFT methods that use analytical basis set. However, special care should be exercised in the choice of the cutoff criteria for the pseudo-atomic orbitals to obtain a good agreement with analytical basis sets approaches.

For the description of the calculated properties, a comparison of PBE1PBE implemented in Gaussian03 and PBE_{SIESTA} allows us to state that the pseudopotential for the platinum derived with the Troullier−Martins procedure and the numerical basis set yield similar results to LanL2DZ and the Pople's basis sets employed in analytical implementations. On the basis of the quality of the results obtained for this type of systems and the computational efficiency of the numerical scheme, SIESTA results an excellent alternative as a computational tool for predicting structure and energetics of platinated systems and their transformations.

Acknowledgment. This work is part of a wider project on Cisplatin chemistry financed by CSIC-UdelaR (Uruguay) through the Research & Development (ELC, 2000–2002) and Young-Researchers (PDD, 2003–2004) programs. Sustained support from PEDECIBA (PNUD-UdelaR) during the 2000–2005 period is also gratefully acknowledged.

References

(1) Jamieson, E. R.; Lippard, S. J. Structure, Recognition, and Processing of Cisplatin-DNA Adducts. *Chem. Rev.* **1999**, *99*, 2467–2496.

(2) (a) Rosenberg, B. Platinum Complexes for the Treatment of Cancer: Why the Search Goes on. In *Cisplatin. Chemistry and Biochemistry of a Leading Anticancer Drug*; Lippert, B., Ed.; Wiley-VCH: Zürich, 1999; pp 3–27. (b) Villani, G.; Le Gac, N. T.; Hoffmann, J.-S. Replication of Platinated DNA and Its Mutagenic Consequences. In *Cisplatin: Chemistry and Biochemistry of a Leading Anticancer Drug*; Lippert, B., Ed.; Wiley-VCH: Zürich, 1999; pp 135–157. (c) Zamble, D. B.; Lippard, S. J. The Response of Cellular Proteins to Cisplatin-Damaged DNA. In *Cisplatin: Chemistry and Biochemistry of a Leading Anticancer Drug*; Lippert, B., Ed.; Wiley-VCH: Zürich, 1999; pp 73–110. (d) Fuertes, M. A.; Alonso, C.; Pérez, J. M. Biochemical modulation of Cisplatin mechanisms of action: enhancement of antitumor activity and circumvention of drug resistance. *Chem. Rev.* **2003**, *103*, 645–662. (e) Wang, D.; Lippard, S. J. Cellular Procesing of Platinum Anticancer Drugs. *Nature Rev.* **2005**, *4*, 307–320.

(3) (a) Basch, H.; Krauss, M.; Stevens, W. J.; Cohen, D. Electronic and geometric structures of $Pt(NH_3)_2^{2+}$, $Pt(NH_3)_2Cl_2$, $Pt(NH_3)_3X$, and $Pt(NH_3)_2XY$ (X, Y = H_2O, OH^-). *Inorg. Chem.* **1985**, *24*, 3313–3317. (b) Kozelka, J.; Savinelli, R.; Berthier, G.; Flament, J. P.; Lavery, R. Force field for platinum binding to adenine. *J. Comput. Chem.* **1993**, *14*, 45–53. (c) Cundari, T. R.; Fu, W.; Moody, E. W.; Slavin., L. L.; Snyder, L. A.; Sommerer, S. O.; Klinckman, T. R. Molecular Mechanics Force Field for Platinum Coordination Complexes. *J. Phys. Chem.* **1996**, *100*, 18057–18064. (d) Pavankumar, P. N. V.; Seetharamulu, S.; Yao, S.; Saxe, J. D.; Reddy, D. G.; Hausheer, F. H. Comprehensive Ab Initio Quantum Mechanical and Molecular Orbital (MO) Analysis of Cisplatin: Structure, Bonding, Charge Density, and Vibrational Frequencies. *J. Comput. Chem.* **1999**, *20*, 365–382. (e) Wysokinski, R.; Michalska, D. The Performance of Different Density Functional Methods in the Calculation of Molecular Structures and Vibrational Spectra of Platinum(II) Antitumor Drugs: Cisplatin and Carboplatin. *J. Comput. Chem.* **2001**, *22*, 901–912. (f) Hofmann, A.; Jaganyi, D.; Munro, O. Q.; Liehr, G.; van Eldik, R. Electronic Tuning of the Lability of Pt(II) Complexes through π-Acceptor Effects. Correlations between Thermodynamic, Kinetic, and Theoretical Parameters. *Inorg. Chem.* **2003**, *42*, 1688–1700. (g) Michalska, D.; Wysokinski, R. Molecular Structure and Bonding in Platinum-Picoline Anticancer Complex: Density Functional Study. *Collect. Czech. Chem. Commun.* **2004**, *69*, 63–72. (h) Wysokinski, R.; Kuduk-Jaworska, J.; Michalska, D. Electronic Structure, Raman and Infrared spectra, and vibrational assignment of Carboplatin. Density functional theory studies. *J. Mol. Struct. (Theochem)* **2006**, *758*, 169–179. (i) Wysokinski, R.; Hernik, K.; Szostak, R.; Michalska, D. Electronic structure and vibrational spectra of cis-diammine-(orotato)platinum(II), a potential Cisplatin analogue: DFT and experimental study. *Chem. Phys.* **2007**, *333*, 37–48.

(4) (a) Carloni, P.; Andreoni, W.; Hutter, J.; Curioni, A.; Giannozzi, P.; Parrinello, M. Structure and bonding in cisplatin and other Pt(II) complexes. *Chem. Phys. Lett.* **1995**, *234*, 50–56. (b) Tornaghi, E.; Andreoni, W.; Carloni, P.; Hutter, J.; Parrinello, M. Carboplatin versus cisplatin: density functional approach to their molecular properties. *Chem. Phys. Lett.* **1995**, *246*, 469–474. (c) Carloni, P.; Andreoni, W. Platinum-Modified Nucleobase Pairs in the Solid State: A Theoretical Study. *J. Phys. Chem.* **1996**, *100*, 17797–17800. (d) Carloni, P.; Sprik, M.; Andreoni, W. Key Steps of the cis-Platin-DNA Interaction: Density Functional Theory-Based Molecular Dynamics Simulations. *J. Phys. Chem. B* **2000**, *104*, 823–835. (e) Spiegel, K.; Rothlisberger, U.; Carloni, P. Cisplatin Binding to DNA Oligomers from Hybrid Car-Parrinello/Molecular Dynamics Simulations. *J. Phys. Chem. B* **2004**, *108*, 2699–2707. (f) Magistrato, A.; Ruggerone, P.; Spiegel, K.; Carloni, P.; Reedjik, J. *J. Phys. Chem. B.* **2006**, *110*, 3604–3613.

(5) (a) Deeth, R. J.; Elding, L. I. Theoretical Modeling of Water Exchange on $[Pd(H_2O)_4]^{2+}$, $[Pt(H_2O)_4]^{2+}$, and *trans*-$[PtCl_2(H_2O)_2]$. *Inorg. Chem.* **1996**, *35*, 5019–5026. (b) Chval, Z.; Sip, M. Pentacoordinated transition states of cisplatin hydrolysis-ab initio study. *J. Mol. Struct. (Theochem)* **2000**, *532*, 59–68. (c) Burda, J. V.; Zeizinger, M.; Sponer, J.; Leszczynski, J. Hydration of cis- and trans-platin: A pseudopotential treatment in the frame of a G3-type theory for platinum complexes. *J. Chem. Phys.* **2000**, *113*, 2224–2232. (d) Bergès, J.; Caillet, J.; Langlet, J.; Kozelka, J. Hydration and 'inverse hydration' of platinum(II) complexes: an analysis using the density functionals PW91 and BLYP. *Chem. Phys. Lett.* **2001**, *344*, 573–577. (e) Zhang, Y.; Guo, Z.; You, X. Hydrolysis Theory for Cisplatin and Its Analogues Based on Density Functional Studies. *J. Am. Chem. Soc.* **2001**, *123*, 9378–9387. (f) Tsipis, A. C.; Sigalas, M. P. Mechanistic aspects of the complete set of hydrolysis and anation reactions of cis- and trans-DDP related to their antitumor activity modeled by an improved ASED-MO approach. *J. Mol. Struct. (Theochem)* **2002**, *584*, 235–248. (g) Costa, L. A. S.; Rocha, W. R.; De Almeida, W. B.; Dos Santos, H. F. The hydrolysis process of the cis-dichloro(ethylendiamine)platinum(II): A theoretical study. *J. Chem. Phys.* **2003**, *118*, 10584–10592. (h) Costa, L. A. S.; Rocha, W. R.; De Almeida, W. B.; Dos Santos, H. F. The solvent effect on the aquation processes of the cis-dichloro(ethylenediammine)platinum(II) using continuum solvation models. *Chem. Phys. Lett.* **2004**, *387*, 182–187. (i) Burda, J. V.; Zeizinger, M.; Leszczynski, J. Activation barriers and rate constants for hydration of platinum and palladium square-planar complexes: An ab initio study. *J. Chem. Phys.* **2004**, *120*, 1253–1262. (j) Robertazzi, A.; Platts, J. A. Hydrogen Bonding, Solvation and Hydrolysis of Cisplatin: A Theoretical Study. *J. Comput. Chem.* **2004**, *25*, 1060–1067. (k) Raber, J.; Zhu, C.; Eriksson, L. A. Activation of anti-cancer drug Cisplatin-is the activated complex fully aquated. *Mol. Phys.* **2004**, *102*, 2537–2544. (l) Zhu, C.; Raber, J.; Eriksson, L. A. Hydrolysis Process of the Second Generation Platinum-Based Anticancer Drug cis-Amminedichlorocyclohexylamineplatinum(II). *J. Phys. Chem. B* **2005**, *109*, 12195–12205. (m) Lau, J. K.-C.; Deubel, D. V. Hydrolysis of the Anticancer Drug Cisplatin: Pitfalls in the Interpretation of Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2006**, *2*, 103–106. (n) Song, T.; Hu, P. Insight into the solvent effect: A density functional theory study of Cisplatin hydrolysis. *J. Chem. Phys.* **2006**, *125*, 091101.

(6) (a) Boudreaux, E. A.; Carsey, T. P. Quasirelativistic MO Calculations on Platinum complexes (Anticancer Drugs) and their Interaction with DNA. *Int. J. Quantum Chem.* **1980**, *18*, 469–479. (b) Basch, H.; Krauss, M.; Stevens, W. J.; Cohen, D. Binding of $Pt(NH_3)_3{}^{2+}$ to nucleic acid bases. *Inorg. Chem.* **1986**, *25*, 684–688. (c) Zilberberg, I. L.; Avdeev, V. I.; Zhidomirov, G. M. Effect of cisplatin binding on guanine in nucleic acid: an ab initio study. *J. Mol. Struct. (Theochem)* **1997**, *418*, 73–81. (d) Pelmenschikov, A.; Zilberberg, I. L.; Leszczynski, J.; Famulari, A.; Sironi, M.; Raimondi, M. cis-$[Pt(NH_3)_2]^{2+}$ coordination to the N7 and O6 sites of a guanine-cytosine pair: disruption of the Watson-Crick H-bonding pattern. *Chem. Phys. Lett.* **1999**, *314*, 496–500. (e) Burda, J. V.; Sponer, J.; Leszczynski, J. The interactions of square platinum(II) complexes with guanine and adenine: a quantum-chemical ab initio study of metalated tautomeric forms. *J. Biol. Inorg. Chem.* **2000**, *5*, 178–188. (g) Burda, J. V.; Sponer, J.; Leszczynski, J. The influence of square planar platinum complexes on DNA base pairing. An *ab initio* DFT study. *Phys. Chem. Chem. Phys.* **2001**, *3*, 4404–4411. (h) Tsipis, A. C.; Katsoulos, G. A. Conformational preferences, rotational barriers and energetics of purine nucleobase rotation and dissociation in square planar platinum(II) antitumour complexes: Structure-activity correlation. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5165–5172. (i) Deubel, D. V. On the Competition of the Purine Bases, Functionalities of Peptide Side Chains, and Protecting Agents for the Coordination Sites of Dicationic Cisplatin Derivatives. *J. Am. Chem. Soc.* **2002**, *124*, 5834–5842. (j) Baik, M.-H.; Friesner, R. A.; Lippard, S. Theoretical Study on the Stability of N-Glycosyl Bonds: Why Does N7-Platination Not Promote Depurination. *J. Am. Chem. Soc.* **2002**, *124*, 4495–4503. (k) Baik, M-H.; Friesner, R. A.; Lippard, S. J. Theoretical Study of Cisplatin Binding to Purine Bases: Why Does Cisplatin Prefer Guanine over Adenine. *J. Am. Chem. Soc.* **2003**, *125*, 14082–14092. (l) Chval, Z.; Sip, M. Transition states of cisplatin binding to guanine and adenine: ab initio reactivity study. *Collect. Czech. Chem. Commun.* **2003**, *68*, 1105–1118. (m) Burda, J. V.; Sponer, J.; Hrabakova, J.; Zeizinger, M.; Leszczynski, J. The Influence of N_7 Guanine Modifications on the Strength of Watson-Crick Base Pairing and Guanine N_1 Acidity: Comparison of Gas-Phase and Condensed-Phase Trends. *J. Phys. Chem. B* **2003**, *107*, 5349–5356. (n) Burda, J. V.; Leszczynski, J. How Strong Can the Bend Be on a DNA Helix from Cisplatin? DFT and MP2 Quantum Chemical Calculations of Cisplatin-Bridged DNA Purine Bases. *Inorg. Chem.* **2003**, *42*, 7162–7172. (o) Deubel, D. V. Factors Governing the Kinetic Competition of Nitrogen and Sulfur Ligands in Cisplatin Binding to Biological Targets. *J. Am. Chem. Soc.* **2004**, *126*, 5999–6004. (p) Jia, M.; Qu, W.; Yang, Z.; Chen, G. Theoretical study on the factors that affect the structure and stability of the adduct of a new platinum anticancer drug with a duplex DNA. *Int. J. Modern Phys. B.* **2005**, *19*, 2939–2949. (q) Raber, J.; Zhu, C.; Eriksson, L. A. Theoretical Study of Cisplatin Binding to DNA: The Importance of Initial Complex Stabilization. *J. Phys. Chem. B.* **2005**, *109*, 11006–11015. (r) Costa, L. A.; Hambley, T. W.; Rocha, W. R.; Almeida, W. B.; Dos Santos, H. F. Kinetics and structural aspects of the cisplatin interactions with guanine: A quantum mechanical description. *Int. J. Quantum Chem.* **2006**, *106*, 2129–2144.

(7) (a) Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. The SIESTA method for ab initio order-N materials simulation. *J. Phys.: Condens. Matter* **2002**, *14*, 2745–2779. (b) Reich, S.; Thomsen, C.; Ordejón, P. Electronic band structure of isolated and bundled carbon nanotubes. *Phys. Rev. B* **2002**, *65*, 155411–155422.

(8) Dans, P. D.; Coitiño, E. L.; Crespo, A.; Estrín, D. A. Unraveling Step by Step the Molecular Choreography Ruling the Sequence-Dependent DNA Structural Changes Promoted by Cisplatin. **2008**, in preparation.

(9) (a) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomeri, 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.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03*, revision B05; Gaussian Inc.: Pittsburgh, PA, 1998. (b) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular

electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(10) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

(11) (a) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868. (b) Adamo, C.; Barone, V. Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters: The mPW and mPW1PW models. *J. Chem. Phys.* **1998**, *108*, 664–675. (c) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *Chem. Phys.* **1993**, *98*, 5648–5652. (d) 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. (e) Burke, K.; Perdew, J. P.; Wang, Y. In *Electronic Density Functional Theory: Recent Progress and New Directions*; Dobson, J. F., Vignale, G., Das, M. P., Eds.; Plenum: New York, 1998; pp 1–395. (f) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(12) (a) Stevens, W.; Basch, H.; Krauss, J. Compact effective potentials and efficient shared-exponent basis sets for the first- and second-row atoms. *J. Chem. Phys.* **1984**, *81*, 6026–6033. (b) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. Relativistic compact effective potentials and efficient, shared-exponent basis sets for the third-, fourth-, and fifth-row atoms. *Can. J. Chem.* **1992**, *70*, 612–630.

(13) (a) Dunning, T. H., Jr.; Hay, P. J. Gaussian basis sets for molecular calculations. In *Modern Theoretical Chemistry*; Schaefer, H. F., III, Ed.; Plenum: New York, 1976; Vol. 3, pp 1–28. (b) Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *J. Chem. Phys.* **1985**, *82*, 270–283. (c) Wadt, W. R.; Hay, P. J. Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi. *J. Chem. Phys.* **1985**, *82*, 284–298. (d) Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.* **1985**, *82*, 299–310.

(14) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724–728.

(15) (a) Crespo, A.; Marti, M. A.; Kalko, S. G.; Morreale, A.; Orozco, M.; Gelpi, J. L.; Luque, F. J.; Estrin, D. A. Theoretical Study of the Truncated Hemoglobin HbN: Exploring the Molecular Basis of the NO Detoxification Mechanism. *J. Am. Chem. Soc.* **2005**, *127*, 4433–4444. (b) Martí, M. A.; Scherlis, D. A.; Doctorovich, F. A.; Ordejón, P.; Estrin, D. A. Modulation of the NO trans effect in heme proteins: implications for the activation of soluble guanylate cyclase. *J. Biol. Inorg. Chem.* **2003**, *8*, 595–600. (c) Martí, M. A.; Capece, L.; Crespo, A.; Doctorovich, F.; Estrin, D. A. Nitric Oxide Interaction with Cytochrome c′ and Its Relevance to Guanylate Cyclase. Why does the Iron Histidine Bond Break. *J. Am. Chem. Soc.* **2005**, *127*, 7721–7728.

(16) Troullier, N.; Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **1991**, *43*, 1993–2006.

(17) Milburn, G. H. W.; Truter, M. R. The crystal structures of cis- and trans-dichlorodiammineplatinum(II). *J. Chem. Soc. A* **1966**, *1*, 1609–1616.

(18) Wing, R. M.; Pjura, P.; Drew, H. R.; Dickerson, R. E. The primary mode of binding of cisplatin to a B-DNA dodecamer: C-G-C-G-A-A-T-T-C-G-C-G. *EMBO J.* **1984**, *3*, 1201–1206.

(19) Hellquist, B.; Bengtsson, L. A.; Holmberg, B.; Hedman, B.; Persson, I.; Elding, L. I. Structures of solvated cations of Palladium(II) and Platinum(II) in dimethyl sulfoxide, acetonitrile and aqueous solution studied by exafs and laxs. *Acta Chem. Scand.* **1991**, *45*, 449–455.

(20) Berners-Price, S. J.; Appleton, T. G. The Chemistry of Cisplatin in Aqueous Solution. In *Platinum-Based Drugs in Cancer Therapy*; Farrell, N. P.; Kelland, L. R., Eds.; Humana Press Inc: Totowa, 2000; pp 3–35..

(21) van der Wijst, T.; Fonseca Guerra, C.; Swart, M.; Bickelhaupt, F. M. Performance of various density functionals for the hydrogen bonds in DNA base pairs. *Chem. Phys. Lett.* **2006**, *426*, 415–421.

CT7002385

JCTC Journal of Chemical Theory and Computation

Development and Validation of the B3LYP/N07D Computational Model for Structural Parameter and Magnetic Tensors of Large Free Radicals

Vincenzo Barone,* Paola Cimino,+ and Emiliano Stendardo

LSDM and INSTM-Village, Dipartimento di Chimica 'Paolo Corradini', Complesso Universitario Monte S. Angelo, via Cintia, I-80126 Napoli, Italy

Received January 31, 2008

Abstract: Extensive calculations on a large set of free radicals containing atoms of the second and third row show that the B3LYP/N07D computational model provides remarkably accurate structural parameters and magnetic tensors at reasonable computational costs. The key of this success is the optimization of core-valence *s* functions for hyperfine coupling constants, while retaining (and even improving) the good performances of the parent 6−31+G(d,p) basis set for valence properties through reoptimization of polarization and diffuse *p* functions.

Introduction

Quantum mechanical treatments of interactions between atomic and molecular systems have provided an invaluable contribution toward a deeper understanding of the interplay between different factors in determining structures, binding energies, and physicochemical properties of noncovalently bonded complexes.[1,2] While very reliable static properties of small and medium size systems can be safely computed by state of the art post-Hartree–Fock methods,[3,4] the situation is more involved for large systems in condensed phases[5] and whenever dynamical aspects cannot be neglected.[6] The development of reliable density functionals (especially hybrid ones),[7–10] mixed discrete/continuum solvent models,[11,12] and implementation of linear scaling computational approaches[13] is allowing the reliable study of large systems of biological and technological relevance.[14] However, the problem of basis set superposition error (BSSE) and of the computation of reliable electric and magnetic properties by basis sets of nonprohibitive dimensions remains open.[15,16] This is even more important in the framework of ab initio dynamics[17,18] where a huge number of different structures (and energy gradients) must be computed to produce a converged trajectory.[19,20] Several recent studies have shown that in the framework of hybrid density functionals and ab initio dynamics, the smallest basis sets allowing semiquantitative evaluations without too large errors connected to basis set incompleteness are split valence sets augmented by diffuse functions.[21,22] Among those, aug-cc-pVDZ[23] and 6–31+G(d,p)[24] models lead to comparable results. From a complementary point of view, the same level of basis sets allows for the computation of reasonable electric and magnetic properties, except for hyperfine coupling constants, which require specialized functions in the core-valence region.[25,26]

Table 1. Basis Functions To Be Added to 6-31G for Obtaining the N07D Basis Set for the B3LYP Functional[a]

	s	*p*	*d*	*d*
H		0.750		
B		0.035	0.343	
C	7.5	0.050	0.820	
N	12.6	0.053	1.015	
O	15.1	0.065	1.190	0.180
F	18.3	0.083	1.370	0.230
Al	3.1	0.015	0.189	
Si	3.6	0.033	0.275	
P	5.5	0.035	0.373	
S	8.0	0.041	0.479	
Cl	8.5	0.048	0.600	0.196

[a] For He, Li, Be, Na, and Mg atoms N07D is identical to 6-31+G(d,p).

All these considerations prompted us to optimize a new polarized split-valence basis set for second- and third-row atoms, which, adding a reduced number of polarization and diffuse functions to the 6–31G set, leads to an optimum

* Corresponding author e-mail: baronev@unina.it.

+ Permanent address: Dipartimento di Scienze Farmaceutiche, Università di Salerno, via Ponte don Melillo, I-84084 Fisciano (Sa), Italy.

10.1021/ct800034c CCC: $40.75
Published on Web 04/09/2008

Table 2. Comparison of Parameters for HF, HCl, H_2O and TEMPO Calculated by Different Basis Sets

	H_2O			HF		HCl		TEMPO	
	OH (Å)	HOH (degrees)	μ (Debye)	HF (Å)	μ (Debye)	HCl (Å)	μ (Debye)	NO (Å)	μ (Debye)
6–31G(d)	0.968	103.7	2.094	0.934	1.860	1.289	1.468	1.286	2.894
aug-cc-pVDZ	0.965	104.7	1.854	0.926	1.803	1.295	1.154	1.283	3.122
6–311+G(2d,2p)	0.961	105.1	1.960	0.922	1.882	1.280	1.174	1.281	3.149
N07D	0.964	104.5	1.846	0.925	1.810	1.291	1.181	1.281	3.162
N07D[a]	0.963	105.3	2.114	0.925	1.986	1.291	1.365	1.281	3.215
exp	0.958[b]	104.5[c]	1.855[d]	0.920[d]	1.826[e]	1.275[f]	1.093[f]		

[a] Without diffuse functions on O, F, and Cl atoms. [b] Reference 52. [c] Reference 53. [d] Reference 54. [e] Reference 55. [f] Reference 56.

Table 3. Theoretical and Experimental Hyperfine Coupling Constants (in Gauss) of B, Be, and Cl Atoms[b]

structure	atom	6–31G(d)	EPR-II	EPR-III	N07D	exp[a]
$BO^{\bullet}$	B	376.0	399.3	384.7	375.9	365.7
$BB^{\bullet\bullet}$	B	9.4	4.7	5.5	8.1	5.4
$BS^{\bullet}$	B	289.6			299.6	292.0
$BH_2^{\bullet}$	B	140.9	133.3	129.1	140.2	127.7
$BH_2O^{\bullet}$	B	−26.7	−28.4	−26.7	−28.0	30.0
$BeH^{\bullet}$	Be	−74.2			−70.4	71.1
$BeOH^{\bullet}$	Be	−106.0			−100.6	94.2
$BeF^{\bullet}$	Be	−111.4			−107.4	104.9
$Cl_2^{-\bullet}$	2Cl	27.1			27.7	38.9
$SiCl_3^{\bullet}$	3Cl	9.2			10.7	12.4
$SiCl_2CH_3^{\bullet}$	2Cl	−7.5			−8.6	10.5
$PCl_2^{\bullet}$	2Cl	−2.3			−2.5	0.4

[a] Data for BO, BB, BS, BH_2, BH_2O, BeH, BeOH, BeF, and Cl_2^- are from ref 44; data for $SiCl_3$, $SiCl_2CH_3$, and PCl_2 are from ref 45. [b] All the theoretical values have been obtained in the present work.

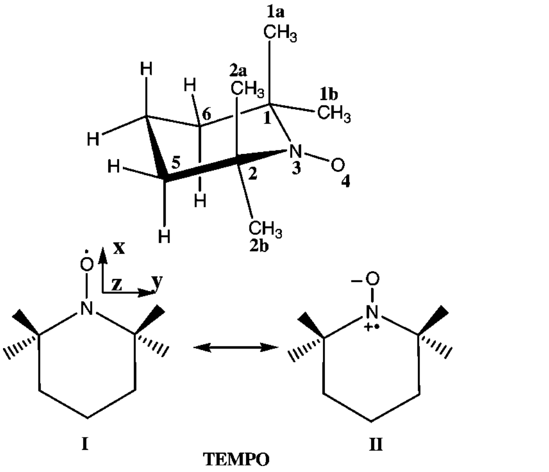


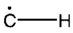
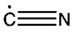
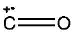
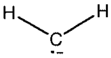
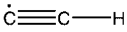
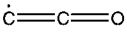
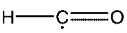
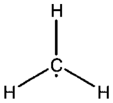
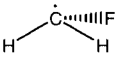
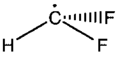
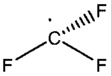
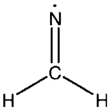
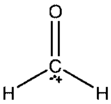
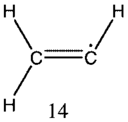
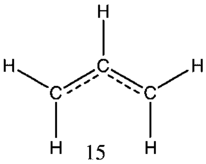
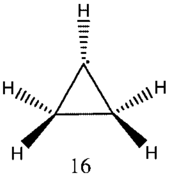
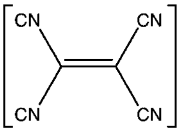
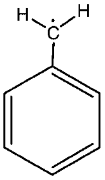
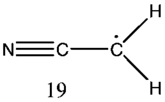
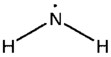
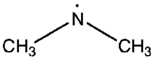
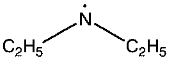
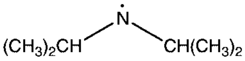
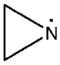
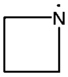
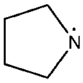
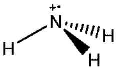
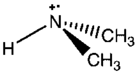
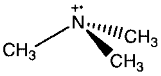
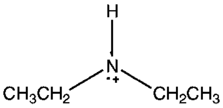
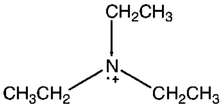
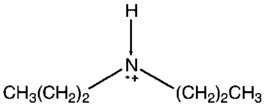
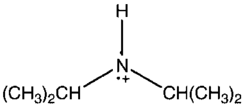
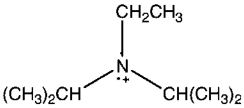
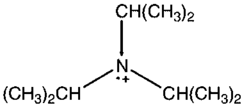
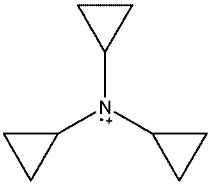
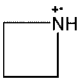
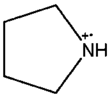
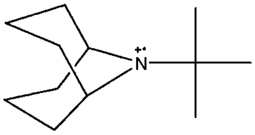
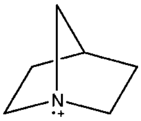
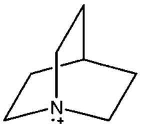
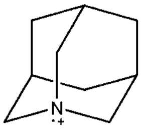
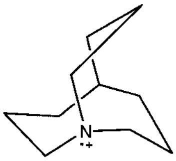
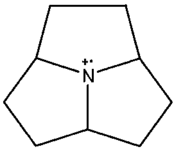
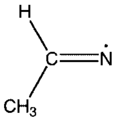
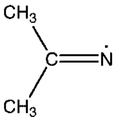
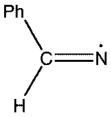
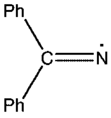
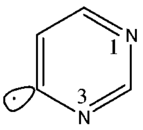
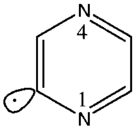
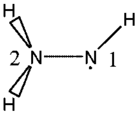
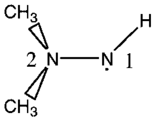
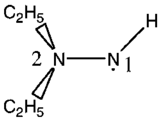
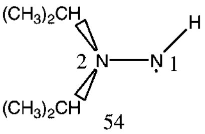
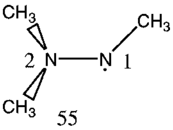
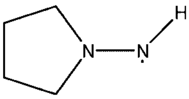
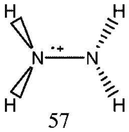
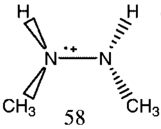
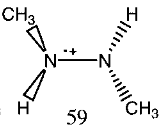
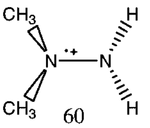
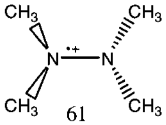
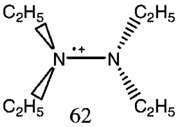
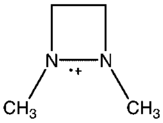
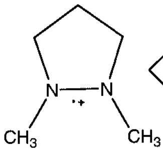
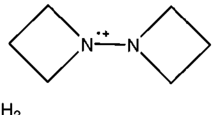
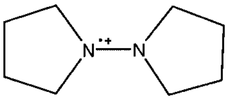
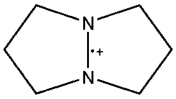
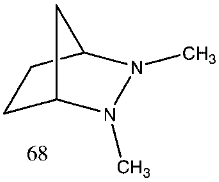
Figure 1. Structure of TEMPO (2,2,6,6-tetramethylpiperidine-*N*-oxyl) radical.

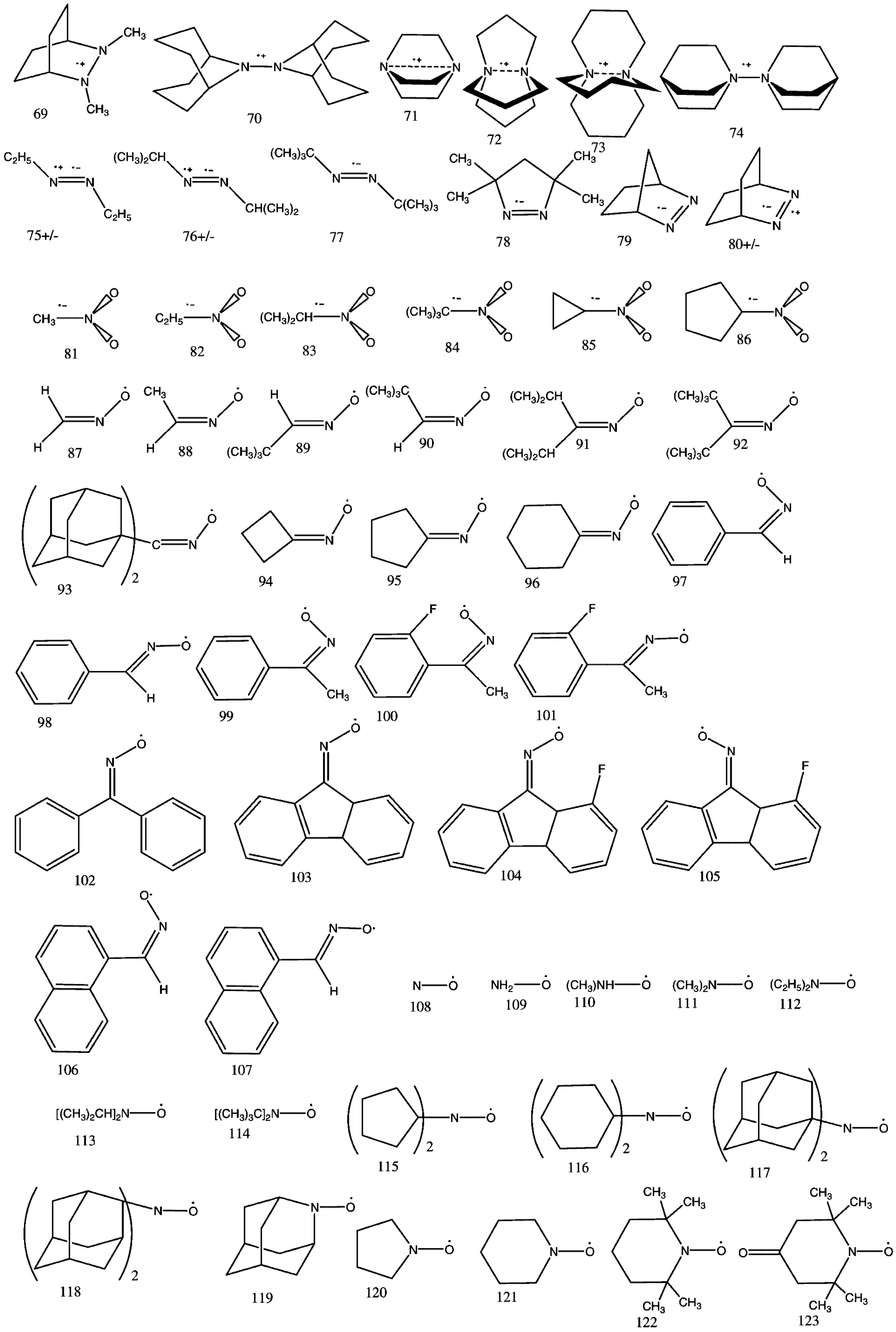
compromise between reliability and computer time. The possible use of different basis sets on different atoms requires basis set balance in order to avoid inaccuracies in the charge distribution of the molecule. For example, adding diffuse functions to split-valence basis sets has a significant effect on the energy even for atoms. Thus, diffuse *p* functions should be added consistently on all non-hydrogen atoms. At the same time, diffuse *s* functions play a negligible role in determining molecular properties and have been neglected for all atoms. The situation is more involved for diffuse *d* functions. Although they have a comparatively lower effect on energies, their role becomes significant for electric properties of electronegative atoms and for some geometrical parameters involving multiple bonds.[27,28] At the same time they adversely affect the basis set superposition error in weak intermolecular interactions. In the present context, diffuse *d* functions have been optimized for all non-hydrogen atoms, but they are systematically used only for some electronegative atoms.

As mentioned above, in the context of magnetic properties, isotropic hyperfine couplings play a peculiar role since their evaluation is quite straightforward, but reliable results can be obtained only by proper inclusion of electron correlation and improved description of core-valence regions.[29–31]

Methods based on the unrestricted Kohn–Sham (UKS) approach to density functional theory (DFT) have revolutionized also this field in the past few years, since some functionals (especially hybrid ones) provide, at least for systems containing only second- and third-row atoms, remarkable results at reasonable costs.[32–37] The basis set issue remains, however, significant.[32] Our experience in developing purposely tailored basis sets indicates that addition of a single core-valence *s* function with an optimized exponent around 10 performs remarkably well.[38,39] The price to be paid for this effective approach is that the basis function must be optimized for each atom and for each different density functional. This is quite disappointing since the remaining part of the basis set can be transferred without modifications among different hybrid density functionals. We have decided, however, that efficiency merits this slight additional effort, and we have optimized semicore *s* functions for several functionals.[40] In the present paper, we will be concerned with the B3LYP functional[41] in view of its widespread use and availability in most computational codes. The functions added to the 6–31G set for all the atoms of the second and third row are shown in Table 1.

The N07D basis set has been already used with success in some structural and dynamic studies both in gas phase


1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68


69
70
71
72
73
74
75+/-
76+/-
77
78
79
80+/-
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
CH3
C2H5
(CH3)2CH
CH(CH3)2
(CH3)3C
C(CH3)3
N
O
H
F
NH2
(CH3)NH
(CH3)2N
(C2H5)2N
[(CH3)2CH]2N
[(CH3)3C]2N
2

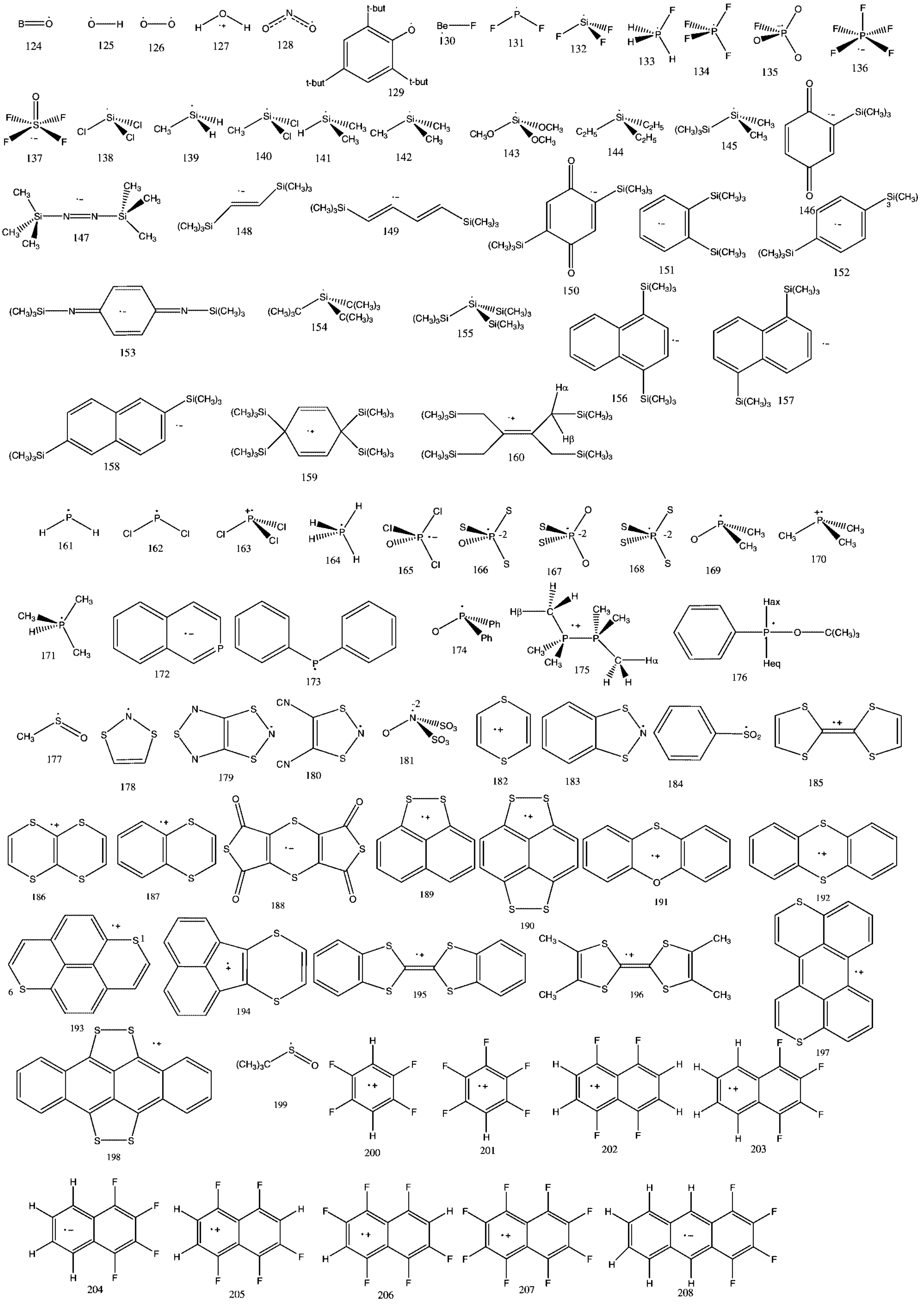


Figure 2. Structures of the radicals studied.

Table 4. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Hydrogen Nuclei of the Radicals Studied[b]

		6–31G(d)	EPR-II	EPR-III	N07D	exp[a]
1	H	−18.2	−17.8	−17.1	−17.0	20.6
4	2H	−17.1	−16.8	−13.8	−14.2	16.0
5	H	21.8	21.9	21.1	21.8	18.0
8	3H	−25.2	−23.8	−23.2	−23.0	25.0
9	2H	−18.3	−17.8	−17.6	−17.4	21.1
10	H	−23.1	−24.5	−25.3	−22.1	22.2
12	H	−78.8	−83.4	−83.0	−79.4	83.2
13	2H	−121.5	−128.2	−128.7	−120.0	132.7
14	Hcis	59.8	63.8	63.8	59.5	68.5
14	Htrans	36.6	40.2	40.2	36.0	34.2
14	H(CH)	14.0	17.1	17.5	13.9	13.3
15	2Hcis	−15.8	−14.9	−14.5	−14.3	13.5
15	2Htrans	−16.5	−15.7	−15.4	−15.0	14.8
15	H(CH)	4.9	4.7	4.5	4.4	4.2
18	H(orto)	−6.0	−5.8	−5.6	−5.9	5.2
18	H(meta)	2.6	2.5	2.4	2.5	1.8
18	H(para)	−6.8	−6.6	−6.5	−6.7	6.2
18	H(CH2)	−18.3	−17.4	−16.8	−17.6	16.3
125	H	−24.2	−24.0	−23.7	−23.4	25.5
127	H	−24.9	−25.6	−25.3	−24.7	26.1
130	2H	−13.6	0.0	0.0	−13.8	11.8
202	4H	−2.0	−2.0	−1.9	−2.0	4.0
203	2H	−5.1	−4.9	−4.9	−5.1	2.4
203	2H	−1.2	−1.2	−1.2	−1.2	0.6
204	2H	−7.3	−6.8	−6.7	−7.1	5.6
204	2H	−2.1	−2.0	−2.0	−2.1	1.4
205	H	−1.1	−1.1	−1.1	−1.1	2.1
205	H	−3.1	−2.9	−2.9	−3.0	4.2
206	2H	−0.1	−0.1	−0.1	−0.1	0.3

[a] Experimental data for 1, 4, 5, 8–10, 12–15, and 18 are from ref 44; for 203–207 are from ref 55. [b] All the theoretical values have been obtained in the present work.

Table 5. Data Analysis for Hydrogen Nuclei[a]

	6–31G(d)	EPR-II	EPR-III	NO7D	exp
		Hydrogen: $N = 29$			
MAD	2.1	1.8	1.9	1.8	
max absolute error	11.3	6.0	6.0	9.0	
average *E*%	25.3%	25.4%	25.4%	25.9%	
max *E*%	113.1%	108.4%	104.6%	113.9%	
R2	0.9946	0.9938	0.9934	0.9943	
intercept	1.4936	0.8670	0.5675	1.0527	
slope	0.9097	0.9666	0.9691	0.9051	
max	121.5	128.2	128.7	120.0	132.7
min	0.1	0.1	0.1	0.1	1.8

[a] MAD (mean absolute deviation in Gauss) = $\Sigma|a_{calc}-a_{exp}|/N$; *E*% (percent error) = $a_{calc} - a_{exp}/a_{exp}$.

Table 6. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Carbon Nuclei

structure	carbon (^{13}C)	6–31G(d)[a]	EPR-III[a]	TZVP[a]	cc-pVQZ[a]	N07D[b]	exp[a]
1	C	16.8	19.0	14.8	9.0	15.5	16.8
2	C	183.9	207.7	214.6	201.6	200.9	209.8
3	C	495.3	569.3	587.8	566.5	555.3	561.3
4	C	33.0	29.4	20.7	16.0	27.3	21.0
5	C	348.2	378.7	390.7	374.3	371.3	362.0
5	C(H)	72.9	81.2	82.9	83.5	78.9	76.0
6	C	16.7	14.2	12.0	8.1	13.3	15.7
6	C(O)	−2.7	−7.1	−8.5	−8.1	−6.5	10.7
7	C	152.2	138.5	142.7	136.8	137.1	134.7
8	C	44.4	28.6	27.0	19.9	28.7	27.0
9	C	62.3	56.4	50.9	40.0	56.3	54.8
10	C	152.8	143.4	146.8	137.6	145.4	148.8
11	C	258.9	264.5	274.0	261.3	266.9	271.6
12	C	−23.6	−24.5	−25.7	−24.4	−24.4	28.9
13	C	−30.1	−33.5	−35.1	−33.2	−33.2	38.9
14	C	121.9	107.7	109.6	101.5	109.5	107.6
14	CH2	−7.0	−4.9	−5.3	−4.0	−4.8	8.6
15	CH	−17.5	−16.0	−16.3	−14.9	−15.9	17.2
15	CH2	28.2	18.3	17.0	13.2	18.6	21.9
16	CH	108.8	93.6	95.5	87.7	95.5	95.9
17	CN	−8.3	−9.2	−9.7	−8.9	−9.1	9.5
18	CH2	32.0	20.4	19.0	15.1	21.0	24.5
18	C1(3)	−14.4	−13.7	−14.1	−13.2	−13.6	14.5

[a] From ref 44. [b] This work.

and in solution (e.g., refs 42 and 43, where it was referred to as N06). In the next section we give just a flavor of its broad performances, whereas the body of the paper is devoted to hyperfine coupling constants, which are one of the main

Table 7. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Nitrogen Nuclei

structure	Nitrogen (^{14}N)	6–31G(d)^{a}	EPR-III^{a}	TZVP^{a}	N07D^{b}	exp^{a}	structure	Nitrogen (^{14}N)	6–31G(d)^{a}	EPR-III^{a}	TZVP^{a}	N07D^{b}	exp^{a}
12	N	10.0	8.2	6.2	11.0	9.2	69	2N	12.0	10.1	9.3	11.7	13.9
19	N	4.8	3.6	2.7	4.8	3.5	70	2N	12.4	10.8	10.0	12.3	13.3
20	N	11.9	10.1	7.8	12.7	10.0	71	2N	16.6	17.1	17.0	17.4	17.0
21	N	−15.2	−12.5	−10.6	−15.7	14.8	72	2N	11.7	10.9	10.1	12.1	14.7
22	N	14.8	12.1	10.4	15.2	14.3	73	2N	29.8	30.8	30.8	31.3	35.9
23	N	15.0	12.5	10.7	15.4	14.3	74	2N	29.9	31.8	31.9	31.1	38.7
24	N	13.0	10.7	8.5	13.7	12.5	75	2N anion	8.1	6.6	5.8	8.0	7.8
25	N	14.8	11.8	9.9	15.3	14.0	75	2N cation	13.9	12.1	11.3	13.2	21.0
26	N	15.4	12.4	10.6	15.7	14.3	76	2N anion	8.6	6.7	6.2	7.6	8.0
27	N	−18.7	−15.0	−13.3	−17.9	19.6	76	2N cation	12.6	10.9	10.2	11.9	20.0
28	N	17.3	14.3	12.8	16.9	19.3	77	2N	8.9	7.2	6.4	8.7	8.2
29	N	18.4	15.6	14.0	18.8	20.7	78	2N	9.5	7.4	7.1	8.1	9.2
30	N	17.1	14.1	12.6	16.5	18.7	79	2N	9.1	6.9	6.5	7.1	8.6
31	N	18.4	15.7	14.2	18.1	20.8	80	2N anion	9.3	7.5	6.8	8.5	8.8
32	N	15.6	13.0	11.7	15.2	18.6	80	2N cation	31.0	33.1	34.0	33.4	31.4
33	N	16.6	13.9	12.6	16.2	18.7	81	N	25.3	23.8	23.2	24.6	25.6
34	N	18.4	15.7	14.4	18.2	20.2	82	N	−24.8	−22.6	−22.3	−19.7	26.0
35	N	18.7	16.1	14.7	18.6	20.2	83	N	24.8	22.0	22.0	26.3	25.4
36	N	13.2	11.4	10.3	13.2	20.1	84	N	27.5	26.2	25.6	22.1	26.6
37	N	17.5	14.4	13.0	17.0	19.1	85	N	22.3	20.2	19.7	21.0	23.8
38	N	17.0	14.0	12.5	16.4	20.0	86	N	24.0	21.6	21.3	21.9	27.0
39	N	17.5	15.1	13.8	17.3	19.5	87	N	30.0	30.1	29.8	31.1	33.3
40	N	26.8	26.7	25.8	28.2	30.2	88	N	31.3	31.7	31.5	32.2	32.5
41	N	22.0	21.4	20.5	22.9	25.1	89	N	28.8	29.1	28.7	30.0	30.5
42	N	19.3	18.3	17.4	19.7	21.6	90	N	30.5	30.5	30.3	31.2	32.2
43	N	17.2	14.8	13.5	17.1	19.2	91	N	30.3	29.8	29.6	30.3	30.7
44	N	20.9	19.5	18.3	22.1	25.0	92	N	30.8	31.1	30.9	32.2	31.3
45	N	10.3	8.3	6.3	11.1	10.2	93	N	30.5	31.0	30.7	31.6	31.1
46	N	10.8	8.4	6.5	11.4	9.6	94	N	28.0	28.3	28.0	29.1	31.6
47	N	10.8	8.5	6.6	11.2	11.3	95	N	29.9	29.7	29.5	30.6	32.2
48	N	11.0	8.6	6.7	11.2	10.0	96	N	30.5	30.8	30.6	31.6	30.7
49	N	29.1	32.4	33.7	32.4	28.0	97	N	31.5	31.9	31.8	32.8	32.6
50	N	30.5	33.6	35.0	33.7	28.0	98	N	28.6	29.1	28.7	30.1	30.0
51	N	12.0	10.3	8.5	12.4	8.8	99	N	31.2	31.6	31.5	32.6	31.6
51	N	12.6	11.1	10.9	10.9	11.7	100	N	31.0	30.9	30.8	31.9	32.0
52	N	11.1	8.7	7.3	10.6	9.6	101	N	30.3	31.0	30.7	35.3	32.0
52	N	12.4	11.7	11.0	12.4	11.5	102	N	31.5	32.1	31.9	32.9	31.5
53	N	11.3	8.7	7.5	10.6	9.6	103	N	30.3	31.0	30.8	32.0	30.9
53	N	11.6	10.6	10.0	11.0	11.1	104	N	30.0	30.4	30.1	31.4	31.1
54	N	9.8	8.7	7.5	10.6	10.0	105	N	32.1	33.1	33.0	34.1	32.6
54	N	11.3	8.7	7.9	9.6	11.7	106	N	32.2	32.6	32.5	33.5	32.4
55	N	12.5	10.3	8.9	12.5	11.7	107	N	29.0	29.6	29.2	30.4	31.0
55	N	11.2	9.9	9.4	10.9	10.5	108	N	6.5	6.6	4.7	8.4	10.6
56	N	10.9	8.5	7.2	10.1	10.6	109	N	−12.8	−11.1	−10.4	−9.8	11.9
56	N	7.9	7.0	6.1	7.5	10.6	110	N	13.3	11.9	11.1	11.9	13.8
57	2N	12.0	9.8	8.8	10.7	11.6	111	N	15.0	14.0	13.0	15.0	15.2
58	2N	12.2	10.1	9.2	11.5	14.7	112	N	14.8	13.6	12.6	14.8	16.7
59	2N	12.5	10.5	9.6	11.8	13.0	113	N	12.3	11.1	10.0	12.2	15.9
60	N	15.7	13.9	12.7	15.7	16.1	114	N	13.9	12.7	11.8	14.2	16.2
60	N	10.8	8.7	8.1	9.7	9.7	115	N	13.3	11.9	10.9	12.9	14.9
61	2N	12.4	10.4	9.5	12.1	13.4	116	N	10.7	9.4	8.3	11.6	14.4
62	2N	11.1	9.3	8.4	10.8	13.2	117	N	13.6	12.3	11.4	13.8	15.2
63	2N	13.2	11.3	10.5	11.1	15.0	118	N	14.1	12.5	11.5	11.3	14.1
64	2N	12.5	10.6	9.7	11.1	15.0	119	N	−17.6	−17.1	−16.7	−21.6	19.8
65	2N	13.2	11.5	10.7	13.1	14.8	120	N	10.1	8.9	7.7	10.7	16.6
66	2N	11.4	9.5	8.6	11.2	12.9	121	N	17.1	16.8	15.9	15.0	16.9
67	2N	16.5	15.5	14.8	17.5	17.6	122	N	14.1	12.9	12.0	14.3	16.2
68	2N	15.0	14.0	13.3	15.7	16.0	123	N	14.2	12.9	12.0	14.3	14.5

^{a} From ref 46. ^{b} This work.

targets of the present development. In this context, a recent systematic study by Hermosilla et al.[44–46] allows for the comparision of the performances of the B3LYP functional for a large set of hyperfine coupling constants employing several basis sets including 6–31G(d),[24] EPR-II,[38] EPR-III,[39] TZVP,[47] and cc-pVQZ.[48] Here we will show that much improved results are consistently obtained by the new N07D basis set with the same functional. As an aside, we have carefully selected a quite large set of experimental data, which represents, in our opinion, a useful benchmark for functional and/or basis set validation.

Computational Details

All the calculations were carried out by the Gaussian03 package[49] using the B3LYP hybrid density functional[41] with the N07D basis set. As mentioned in the Introduction, this basis set was obtained adding to a double-ζ description of

Table 8. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Oxygen Nuclei

structure	Oxygen (^{17}O)	6–31G(d)[a]	EPR-III[a]	TZVP[a]	cc-pVQZ[a]	N07d[b]	exp[a]
3	O	11.1	9.9	10.2	9.5	10.4	6.6
7	O	−11.6	−12.5	−10.5	−8.1	−14.1	15.1
87	O	−18.1	−17.9	−14.1	−9.8	−21.2	22.8
123	O					−18.0	19.3
124	O	−0.2	−4.2	−4.1	−4.6	−3.5	5.0
125	O	−18.3	−15.7	−8.6	−1.5	−22.0	18.3
126	2O	−15.1	−13.9	−10.3	−5.8	−17.3	19.6
127	O	−31.0	−23.1	−16.0	−8.5	−29.3	29.7
128	2O	−14.5	−19.4	−16.9	−14.6	−20.3	21.8
129	O					−11.2	10.2
188	O			−2.2	−1.4	−3.5	3.6
199	O					−15.0	15.5

[a] From ref 44. [b] This work.

Table 9. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Fluorine Nuclei

structure	Fluorine (^{19}F)	6–31G(d)[a]	EPR-III[a]	TZVP[a]	cc-pVQZ[a]	N07D[b]	exp[a]
9	F	−73.2	−52.2	−51.3	−44.7	−61.3	64.3
10	2F	−71.8	−77.2	−72.9	−62.4	−79.0	84.2
11	3F	133.9	138.3	133.4	125.6	134.5	142.4
130	F	71.1	92.0	84.8	91.4	86.2	81.7
131	F	47.0	31.0	29.5	19.6	31.5	32.6
134	2Feq	59.8	52.7	51.8	50.7	51.0	60.0
135	F	−14.1	−11.2	−12.4	−10.7	−12.6	8.0
136	4Feq	211.5	187.0	182.2	172.7	186.1	206.6
200	F		21.9			24.7	25.8
201	F1		−6.2			−6.7	4.8
201	F2,6		21.4			25.8	25.8
201	F3,5		22.8			24.2	25.8
202	F		15.9			17.7	16.2
203	F1,4		18.8			20.7	19.5
203	F2,3		4.0			4.5	6.5
204	F1,4		6.4			7.8	6.1
204	F2,3		2.7			2.7	2.1
205	F1		13.6			15.1	16.1
205	F3		6.0			6.9	7.1
205	F4		17.3			19.3	16.1
205	F5		17.9			19.8	16.8
205	F8		14.1			16.1	16.1
206	F1,5		15.1			16.7	17.9
206	F3,7		8.7			9.8	10.3
206	F4,8		15.5			17.3	17.9
207	F1,4,5,8		16.4			18.2	19.0
207	F2,3,6,7		3.3			3.8	4.8
208	F1,4		4.2			4.8	4.5
208	F2,3		2.5			2.5	2.3

[a] From ref 44 for 9, 10, 11, 130, 131; from ref 45 for 134, 135, 136, and from ref 55 for 200–208. [b] This work.

valence orbitals single sets of optimized core-valence *s* (on all atoms except H), diffuse *p* (on all atoms except H), polarization (on all atoms), and diffuse *d* (on O, F, Cl atoms) functions (Table 1). The inner electrons of second- and third-row atoms were described by the 6G basis set.[24]

Geometry optimizations and evaluations of harmonic frequencies have been performed in the gas phase using analytical gradients and Hessians. Nuclear hyperfine tensors have been computed following well-defined procedures described in recent literature.[32,50]

The hyperfine coupling tensor ($\boldsymbol{A}_X$), which describes the interaction between the electronic spin density and the nuclear magnetic momentum of nucleus *X*, can be split into three terms: $\boldsymbol{A}_X = a_X\boldsymbol{I}_3 + \mathbf{T}_X + \boldsymbol{\Lambda}_X$, where $\boldsymbol{I}_3$ is the 3 × 3 unit matrix. The first term (a_X), usually referred to as the Fermi-contact interaction, is an isotropic contribution, also known as a hyperfine coupling constant (hcc), and is related to the spin density at the corresponding nucleus *X*. The second contribution ($\boldsymbol{T}_X$) is anisotropic and can be derived from the classical expression of interacting dipoles. The last term, $\boldsymbol{\Lambda}_X$, is due to second-order spin–orbit coupling and can be determined by methods similar to those used for the *g*-tensor.[51] In the present case, because of the strong localization of spin density on the studied atoms and of their small spin–orbit coupling constants, its contribution can be safely neglected and will not be discussed in the following. Of course, upon complete averaging by rotational motions, only the isotropic part survives.

Results and Discussion

The N07D basis set has been assessed by comparison with some standard basis sets for a number of properties: a) geometrical parameters, b) dipole moments, and c) hyperfine coupling constants. Our results are collected in Tables 2−16

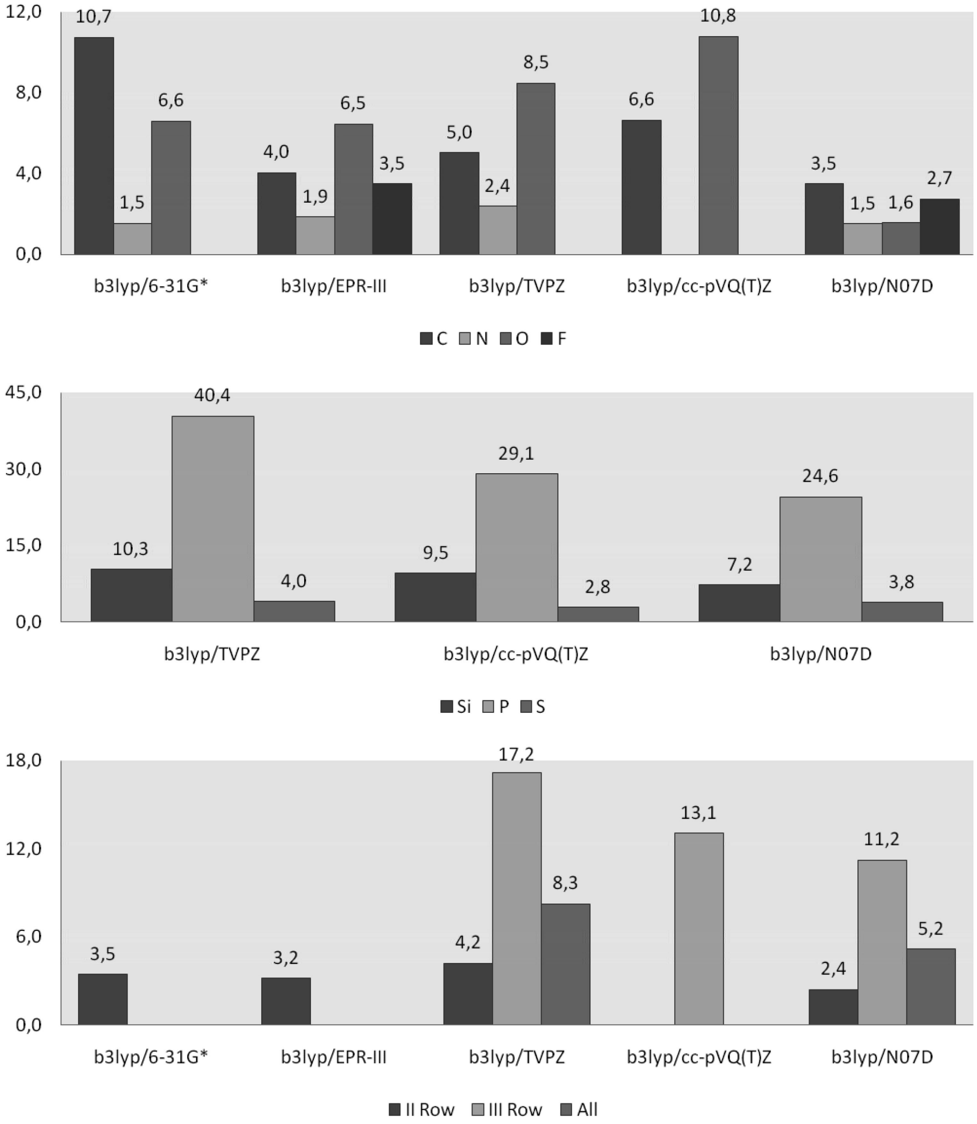


Figure 3. MADs (in Gauss) for nuclei of second and third rows. MAD (mean absolute deviation = $\Sigma|a_{\mathrm{calc}}-a_{\mathrm{exp}}|/N$).

and compared with other available theoretical and experimental results.

Before discussing the results in some detail, let us point out that for medium size basis sets (e.g., 6–31G(d,p), cc-pVDZ, N07D) the number of components of *d* functions plays an important role in obtaining accurate hyperfine coupling constants: in particular it is mandatory to use the redundant set of six *d* functions (which is the standard for 6–31G-like basis sets) because the additional *s* function implicitly added when using a 6d set plays a non-negligible role in completing the *s* space. Although this is not the case for larger basis sets (essentially equivalent results are obtained by the EPR-III basis set using 5d or 6d functions),[32,46] we think that this is a modest price to be paid for the much reduced cost of N07D wrt EPR-III. From another point of view, diffuse polarization functions play a significant role for electric properties, which are, in turn, related to noncovalent interactions. As a compromise between accuracy and cost, we decided, on the ground of test computations (some of which are discussed in the next section) to add diffuse *d* functions only on O, F, and Cl atoms.

Geometric Parameters and Electric Properties. The performances of the N07D basis set are generally comparable to those of aug-cc-pVDZ, with increased computational efficiency. For purposes of illustration, we report in Table 2 some significant parameters of H_2O, HF, HCl, and of the nitroxide radical TEMPO (Figure 1).[52–56] While structural parameters are generally satisfactory, irrespective of the presence of diffuse functions on O, F, and Cl atoms, dipole moments are significantly improved by the addition of diffuse polarization functions, reaching quantitative agreement with experiment. Thus the range of application of the B3LYP/N07D model is significantly enlarged by addition of diffuse polarization functions on electronegative atoms.

Hyperfine Coupling Constant. A variety of molecules containing hydrogen and atoms from the second- and third-row of the periodic table have been studied. We have taken 199 radicals (for a total of 221 hcc's) considered by Hermosilla and co-workers,[44–46] together with 9 additional radicals containing fluorine atoms.[57] The selected set (shown in Figure) includes neutral, cationic, anionic, doublet, triplet, quartet, localized, and conjugated radicals.

Table 10. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Silicon Nuclei

structure	Silicon (^{29}Si)	TZVP[a]	cc-pVQZ[a]	N07D[b]	exp[a]
132	Si	−457.1	−456.9	497.9	498.0
138	Si	−391.8	−404.0	426.8	416.0
139	Si	−162.7	−161.4	164.6	181.0
140	Si	−275.7	−280.7	290.8	295.0
141	Si	−161.3	−160.2	164.2	183.0
142	Si	−159.8	−158.7	162.6	181.0
143	Si	−303.0	−305.8	311.0	339.0
144	Si	−145.5	−144.1	147.6	170.0
145	Si	−121.1	−125.3	122.9	137.0
146	Si	1.5	1.6[a]	−1.6	1.5
147	Si	6.8	7.5[a]	−7.1	7.0
148	2Si	6.6	7.9[a]	−5.6	6.7
149	2Si	6.1	6.6[a]	−5.6	5.7
150	Si	1.3	1.2[a]	−1.5	1.5
151	2Si	5.0	5.8[a]	−3.8	4.5
152	Si	5.8	6.7[a]	−4.9	6.2
153	2Si	4.2	4.2[a]	−4.3	3.9
154	Si	−140.4	−143.0[a]	137.5	163.0
155	Si	−62.8	−63.3[a]	53.9	64.0
155	3Si	4.0	3.4[a]	−4.9	7.1
156	2Si	5.2	5.4[a]	−5.1	4.6
157	2Si	3.9	4.0[a]	−3.9	3.5
158	2Si	2.9	3.3[a]	−3.5	2.7
159	4Si	−18.7	−19.1[a]	19.8	20.9
160	4Si	−9.8	−10.0[a]	10.2	12.5

[a] From ref 45. [b] This work.

Table 11. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Phosphorus Nuclei

structure	Phosphorus (^{31}P)	TZVP[a]	cc-pVQZ[a]	N07D[bb]	exp[a]
131	P	96.9	72.1	65.1	84.8
133	P	702.6	701.9	728.4	721.3
134	P	1203.3	1241.8	1305.8	1330.0
135	P	−47.8	−41.9	−46.9	39.1
136	P	1262.3	1290.5	1427.6	1328.2
161	P	76.9	60.2	64.8	77.4
162	P	61.8	56.4	46.0	68.3
163	P	720.6	761.1	784.9	833.5
164	P	479.5	487.8	536.7	519.3
165	P	1248.5	1314.2	1312.9	1371.0
166	P	−12.5	−8.7	−15.7	13.5
167	P	−12.7	−9.2	−17.7	16.8
168	P	−12.5	−8.3	−16.5	14.7
169	P	302.2	319.5	317.4	375.0
170	P	322.1	322.2	329.1	388.9
171	P	469.1	475.2	495.1	484.0
172	P	22.4	19.8	18.5	23.6
173	P	61.9	70.7	72.2	78.7
174	P	297.3	312.9	318.5	361.6
175	2P	439.5	459.2	479.9	482.0
176	P	508.7	530.4	547.6	557.0

[a] From ref 45. [b] This work.

Table 12. Theoretical and Experimental Hyperfine Coupling Constants (Gauss) of Sulfur Nuclei

structure	Sulfur (^{33}S)	TZVP	cc-pVQZ	N07D[b]	exp
137	S	314.3	335.7	329.3	362.6
177	S	7.8	7.5	6.6	8.0
178	2S	3.4	3.2	3.2	4.2
179	2S	2.2	2.1	2.1	3.3
180	2S	3.4	3.1	3.2	4.4
181	2S	2.1	2.0	3.4	1.3
182	2S	7.9	7.1	6.9	9.8
183	2S	3.1	2.7	2.8	3.9
184	S	61.2	65.8	64.6	83.2
185	4S	3.5	3.2	3.0	4.3
186	4S	3.5	3.1	3.0	4.2
187	2S	7.6	6.4	6.5	9.4
187	2S	−0.8	−0.8[a]	−0.8	0.8
188	S2,6	−1.0	−1.1[a]	−0.8	1.4
189	2S	5.2	5.9[a]	4.2	7.2
190	S1,2,5,6	3.6	4.2[a]	3.1	4.4
191	S	8.4	9.3[a]	7.2	11.9
192	2S	7.1	7.8[a]	6.0	9.2
193	2S	4.4	5.0[a]	3.9	5.3
194	S	6.2	6.9[a]	5.4	8.3
195	4S	3.3	3.7[a]	2.8	4.1
196	4S	3.2	3.6[a]	2.7	4.0
197	2S	4.1	4.5[a]	3.5	4.6
198	4S	2.4	2.9[a]	2.1	3.4

[a] From ref 45. [b] This work.

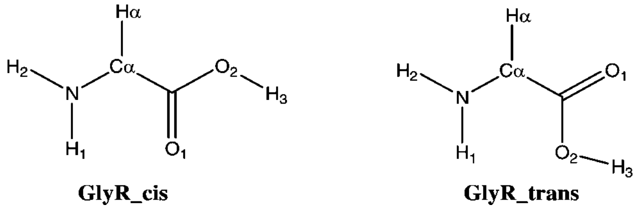


Figure 4. Structures of the glycine radical.

Before considering detailed results, we point out that the reduced number of experimental data available for Be, B, and Cl does not allow for a significant statistical analysis: the corresponding results are thus collected in Table 3 for purposes of illustration only: it is quite apparent that the N07D basis set delivers in all cases reasonable results.

For all the other atoms, we report the number of data (N), mean absolute deviation (MAD), data range, average absolute error, and mean percent error (MPE) between calculated and experimental values. Next we give the correlation coefficient (R^2), slope, and intercept of the least-squares line. The MAD and MPE only consider the absolute value, so that all deviations are converted to positive numbers, added, and then averaged. Since the absolute errors increase, of course, with the range spanned by the corresponding hcc's, the error expressed in a percentage basis would seem coherent and intuitive: however, this procedure gives rise to serious difficulties with hcc's that are very small or close to zero. In such circumstances, regression analysis represents, in our opinion, the simplest and most useful approach for an unbiased comparison between large sets of computed and experimental hcc's.

Hydrogen atoms require some specific considerations in view of the lack of inner shells and of the overwhelming role of small hcc's in an unbiased statistics. We have thus selected a specific set of data in which the presence of σ radicals (characterized by large hcc's) has been overemphasized. The results collected in Tables 4 and 5 show that different basis sets are nearly equivalent in this connection leading to a percent error around 10%, which is close to that of second-row atoms.

We analyze the data for the other atoms in separate parts. In the first one, we consider each nucleus separately; next, atoms belonging to the same row of the periodic table are grouped together, and, finally, all the atoms are taken as a single set. We compare our results (Tables 6−12) with both experimental data and theoretical ones making explicit reference, in the latter case, to the B3LYP results with

Table 13. Data Analysis for Second Row Nuclei[a]

	6–31G(d)	EPR-III	TVPZ	cc-pVQZ	NO7D	exp
Carbon: $N = 23$						
MAD	10.7	4.0	5.0	6.6	3.5	
max absolute error	66.0	16.7	28.7	14.8	9.3	
average *E*%	18.6%	10.9%	9.3%	18.7%	10.5%	
max *E*%	74.8%	42.7%	38.0%	53.2%	43.9%	
R2	0.9915	0.9988	0.9991	0.9983	0.9991	
intercept	7.5938	−1.7012	−3.6841	−6.2305	−1.1153	
slope	0.9033	1.0182	1.0562	1.0196	0.9987	
max	495.3	569.3	587.8	566.5	555.3	561.3
min	2.7	4.9	5.3	4.0	4.8	8.6
Nitrogen: $N = 105$						
MAD	1.5	1.9	2.4		1.5	
max absolute error	6.5	7.7	8.9		5.9	
average *E*%	7.9%	10.3%	13.5%		8.1%	
max *E*%	39.2%	46.4%	55.7%		35.4%	
R2	0.9725	0.9751	0.9740		0.9771	
intercept	−2.2515	−4.6346	−6.3376		−3.8563	
slope	1.0314	1.1170	1.1639		1.1270	
max	32.2	33.1	33.0		35.3	33.3
min	6.5	6.6	4.7		8.4	10.6
Oxygen: $N = 12$						
MAD	6.6	6.5	8.5	10.8	1.6	
max absolute error	19.3	19.3	19.3	21.2	3.8	
average *E*%	33.6%	22.7%	38.8%	52.7%	16.4%	
max *E*%	96.0%	50.0%	54.5%	91.8%	57.7%	
R2	0.7982	0.9190	0.7383	0.2041	0.9359	
intercept	−1.1787	2.7574	2.8204	3.7203	1.0446	
slope	0.9311	0.6806	0.4738	0.2128	0.9246	
max	31.0	23.1	16.9	14.6	29.3	29.7
min	0.2	4.2	2.2	1.4	3.5	3.6
Fluorine: $N = 29$						
MAD		3.5			2.7	
max absolute error		19.6			20.5	
average *E*%		14.6%			12.4%	
max *E*%		40.0%			20.3%	
R2		0.9914			0.9956	
intercept		0.0117			1.6564	
slope		0.9320			0.9145	
max		187.0			186.1	206.6
min		2.5			2.5	2.1

[a] MAD (mean absolute deviation in Gauss) = $\Sigma|a_{calc}-a_{expl}|/N$(total nuclei); *E*% (percent error) = $a_{calc} - a_{exp}/a_{exp}$.

different basis sets reported in refs 44−46 and to some new EPR-III computations for aromatic radicals containing fluorine atoms. Tables 13−16 collect the results of the different statistical analyses, and Figure 3 sketches the MADs for each atom.

For the radicals containing carbon and oxygen atoms we compare our data with the theoretical ones calculated with four different basis sets [6–31G(d), EPR-III, TVPZ, and cc-pVQZ], whereas for nitrogen we have three basis sets [6–31G(d), EPR-III, and TVPZ] and for fluorine just one [EPR-III]. In general, all DFT methods yield a_X values close to the experimental ones, and the best results are consistently delivered by the N07D basis set. The performances of the different basis sets are compared in Figure 3 (MADs) and in Tables 13−15 (statistical analysis data). The N07D results for carbon and oxygen atoms are much better than those delivered by other (even significantly larger) basis sets both in terms of MADs (3.5 and 1.6 Gauss, respectively) and closeness of the slope of the linear regression to the theoretical value of 1.0 (0.999 and 0.925, respectively). The comparison for fluorine atoms is restricted to the EPR-III basis set due to the lack of other results: also in this case the N07D MAD is significantly better (2.7 vs 3.5 Gauss). As previously pointed out,[44] the 6–31G(d) basis set delivers particularly good results for nitrogen (albeit still inferior to their N07D counterparts), but this good behavior does not extend to other atoms, and, as said before, the geometries and interaction energies delivered by this basis set are not fully satisfactory. The B3LYP/N07D model gives by far the lowest MAD (Figure 3) for the whole set of C, N, O, F hcc's (2.4 Gauss). Moreover, the results of a linear regression for all nuclei of the second row (155), summarized in Table 8, shows that with N07D R^2 is higher (0.997) than with all the other basis sets, the slope is close to 1.0 (0.991), and the intercept is quite small (−0.873 Gauss).

For the third-row atoms, previous results have been obtained only with the TVPZ and cc-pVQZ basis sets, due to the lack of purposely tailored (e.g., EPR-III) basis sets. Except for the sulfur atom, for which the lowest MAD value is obtained with the cc-pVQZ basis set (2.8 vs 3.8 Gauss), N07D shows the best results (Figure 3). For all three atoms the R^2 value is higher than 0.99, and the slopes are considerably improved by using the N07D basis set and show values close to unit (Tables 7 and 8). The complete regression analysis performed for all third atoms indicates that the results are very satisfactory, indeed, the R^2 is 0.9958, the slope is 0.9956, and the MAD is the lowest among the available basis sets (Table 15). In summary, we can conclude

Table 14. Data Analysis for Third-Row Nuclei[a]

	TVPZ	cc-pVQZ	NO7D	exp
Silicon: $N = 25$				
MAD	10.3	9.5	7.2	
max absolute error	40.9	41.1	28.0	
average E%	10.6%	13.8%	11.1%	
max E%	43.7%	52.1%	31.7%	
R2	0.9988	0.9978	0.9955	
intercept	−1.0288	−0.9419	−4.0187	
slope	0.9166	0.9257	0.9803	
max	457.1	456.9	497.9	498.0
min	1.3	1.2	1.5	1.5
Phosphorus: $N = 21$				
MAD	40.4	29.1	24.6	
max absolute error	126.7	88.2	99.4	
average E%	11.5%	14.3%	10.8%	
max E%	24.4%	45.2%	32.7%	
R2	0.9968	0.9983	0.9942	
intercept	−3.3470	−10.7315	−13.1618	
slope	0.9196	0.9586	1.0060	
max	1262.3	1314.2	1427.6	1371.0
min	12.5	8.3	15.7	13.5
Sulfur: $N = 25$				
MAD	4.0	2.8	3.8	
max absolute error	48.3	26.9	33.3	
average E%	21.8%	19.2%	35.7%	
max E%	61.5%	53.8%	164.9%	
R2	0.9988	0.9989	0.9988	
intercept	−0.6699	−0.9042	−1.3765	
slope	0.8625	0.8952	0.9058	
max	314.3	335.7	329.3	362.6
min	0.8	0.8	0.8	0.8

[a] MAD (mean absolute deviation in Gauss) = $\Sigma|a_{\text{calc}}-a_{\text{exp}}|/N$(total nuclei); E% (percent error) = $a_{\text{calc}} - a_{\text{exp}}/a_{\text{exp}}$.

Table 15. Data Analysis for Second- and Third-Row and All Nuclei

	6–31G(d)	EPR-III	TVPZ	cc-pVQZ	NO7D	exp
II Row Atoms: $N = 155$						
MAD	3.5	3.2	4.2		2.4	
max absolute error	66.0	19.6	28.7		20.5	
average E%	12.9%	15.1%	20.4%		11.6%	
max E%	96.0%	50.0%	55.7%		57.7%	
R2	0.9903	0.9967	0.9955		0.9973	
intercept	1.1825	−2.6255	−4.3727		−0.8730	
slope	0.9364	1.0228	1.0544		0.9905	
max	495.3	569.3	587.8		555.3	561.3
min	0.2	3.6	2.7		3.5	3.5
III Row Atoms: $N = 71$						
MAD			17.2	13.1	11.2	
max absolute error			126.7	88.2	99.4	
average E%			14.7%	15.8%	19.1%	
max E%			61.5%	53.8%	164.9%	
R2			0.9979	0.9986	0.9958	
intercept			−1.7333	−3.7425	−5.8020	
slope			0.9171	0.9491	0.9956	
max			1262.3	1314.2	1427.6	1371.0
min			0.8	0.8	0.8	0.8
All Atoms: $N = 226$						
MAD			8.3		5.2	
max absolute error			126.7		99.4	
average E%			18.6%		14.0%	
max E%			61.5%		164.9%	
R2			0.9966		0.9963	
intercept			−1.1502		−2.1904	
slope			0.9246		0.9913	
max			1262.3		1427.6	1371.0
min			0.8		0.8	0.8

[a] MAD (mean absolute deviation in Gauss) = $\Sigma|a_{\text{calc}}-a_{\text{exp}}|/N$(total nuclei); E% (percent error) = $a_{\text{calc}} - a_{\text{exp}}/a_{\text{exp}}$.

that the B3LYP model couples computational efficiency and reliability for radicals involving atoms of the second and third row.

The performances of the B3LYP/N07D model for a typical problem involving at the same time stereoelectronic, vibrational, and environmental effects can be judged by the results

Table 16. Hyperfine Coupling Constants (in Gauss) of Glycine Radical

	exp	EPR-II average		N07D minimum		N07D average	
	pH: 1–10	cis	trans	cis	trans	cis	trans
N	6.4	5.5	5.5	4.1	4.1	6.3	6.3
H1	5.6	−5.4	−5.6	−7.7	−7.5	−5.3	−5.1
H2	5.6	−5.3	−5.4	−7.9	−7.7	−5.3	−5.1
Hα	11.8	−11.9	−11.7	−14.0	−14.1	−11.6	−11.7

reported in Table 16 for the glycine radical (GlyR, Figure 4) in aqueous solution.[43,58,59] Since the hcc's computed for the minimum energy structure in vacuum are significantly tuned by both intramolecular vibrations and by solvent librations, the reported results are obtained by averaging over 100 frames extracted at regular time steps from the ab initio dynamics described in ref 43. From a general point of view, all the computations provide, as expected, positive values for the C^{α} and N hcc's and negative values for the hydrogen atoms. Moreover, dynamical effects reduce the differences between the pairs H_1, H_2, and C^{α}, H^{α}. Polar solvents increase delocalization along the GlyR backbone, due to an increased importance of ionic resonance structures characterized by double N−C and C−C′ bonds and to the concomitant reduction of H^{α} hcc and of the pyramidalization of the aminic moiety. This last structural effect induces both a significant reduction of the H_2 hcc and an increased delocalization of the molecular orbital formally containing the unpaired electron, with the consequent reduction of the C^{α} and H^{α} hcc's. After averaging by MD in aqueous solution, the computed values are in general good agreement with experiment. It is, however, quite apparent that hydrogen atoms are described in a nearly equivalent way by the EPR-II and N07D basis set, whereas the nitrogen hcc is significantly improved by the new basis set, which is able to deliver quantitative agreement with experiment.

Concluding Remarks

Some hybrid functionals such as the popular B3LYP model are able to treat in a balanced way the differential spin polarization of different shells, thus providing a good description of the magnetic properties of many classes of compounds. Optimization of core-valence, diffuse, and polarization functions in a medium size basis set further extends the reliability of the computational model for both structural and magnetic properties. For all second-row atoms the B3LYP/N07D model couples quantitative agreement with experimental data and predictive power. The situation is slightly worse for third-row atoms, where recourse to regression analysis could prove valuable.

All in all, the B3LYP/N07D results seem accurate enough to allow for quantitative studies, especially taking into account that the same model and basis set can be used for different properties and for second- and third-row atoms. Furthermore, the availability of effective discrete/continuum solvent models and of different dynamical approaches, together with the reduced dimensions of the N07D basis set, allows for the performing of comprehensive analyses aimed at evaluating the roles of stereoelectronic, vibrational, and environmental effects in determining the overall properties of large flexible radicals of current biological and/or technological interest.

References

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

(2) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247.

(3) Christiansen, O. *Theor. Chem. Acc.* **2006**, *116*, 106.

(4) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(5) Barone, V.; Polimeno, A. *Chem. Soc. Rev.* **2007**, *36*, 1724.

(6) (a) Barone, V. *J. Chem. Phys.* **1994**, *101*, 10666. (b) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.

(7) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

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

(9) Zhao, Y.; Shultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput* **2006**, *2*, 364.

(10) Sato, S.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 234114.

(11) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(12) Brancato, G.; Barone, V.; Rega, N. *Theor. Chem. Acc.* **2007**, *117*, 1001.

(13) Scuseria, G. E. *J. Phys. Chem. A* **1999**, *103*, 4782.

(14) Benzi, C.; Improta, R.; Scalmani, G. *J. Comput. Chem.* **2002**, *23*, 341.

(15) Check, C. E.; Faust, T. O.; Bailey, J. M.; Wright, B. J.; Gilbert, T. M.; Sunderlin, L. S. *J. Phys. Chem. A* **2001**, *105*, 8111.

(16) Engels, B.; Peyerimhoff, S. D. *J. Phys. B* **1988**, *21*, 3459.

(17) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(18) Schlegel, H. B.; Millam, J. M.; Iyengar, S. S.; Voth, G. A.; Daniels, D. A.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2001**, *114*, 9758.

(19) Crescenzi, O.; Pavone, M.; de Angelis, F.; Barone, V. *J. Phys. Chem. B* **2005**, *109*, 445.

(20) Pavone, M.; Cimino, P.; De Angelis, F.; Barone, V. *J. Am. Chem. Soc.* **2006**, *128*, 4338.

(21) Iyengar, S. S.; Frisch, M. J. *J. Chem. Phys.* **2004**, *121*, 5061.

(22) Improta, R.; Barone, V.; Kudin, K. N.; Scuseria, G. E. *J. Am. Chem. Soc.* **2001**, *123*, 3311.

(23) Peterson, K. A.; Kendall, R. A.; Dunning, T. H. *J. Chem. Phys.* **1993**, *99*, 9790.

(24) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(25) Chipman, D. M. *Phys. Rev. A* **1989**, *39*, 475.

(26) Barone, V. *Theor. Chim. Acta* **1995**, *91*, 113.

(27) (a) Halls, M. D.; Schlegel, H. B. *J. Chem. Phys.* **1998**, *109*, 10587. (b) Halls, M. D.; Velkovski, J.; Schlegal, H. B. *Theor. Chim. Acc.* **2001**, *105*, 413.

(28) Barone, V. *J. Phys. Chem. A* **2004**, *108*, 4146.

(29) Huang, M. B.; Suter, H. U.; Engels, B.; Peyerimhoff, S. D.; Lunell, S. *J. Phys. Chem.* **1995**, *99*, 9724.

(30) Chipman, D. M. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Ronghoff, S., Ed.; Kluwer: Amsterdam, 1995; p 109.

(31) Al Derzi, A. R.; Fau, S.; Bartlett, R. J. *J. Phys. Chem. A* **2003**, *107*, 6656.

(32) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231.

(33) Barone, V.; Adamo, C.; Russo, N. *Chem. Phys. Lett.* **1993**, *212*, 5.

(34) Barone, V. *J. Phys. Chem.* **1995**, *99*, 11659.

(35) Schoneborn, J. C.; Neese, F. W.; Thiel, *J. Am. Chem. Soc.* **2005**, *127*, 5840.

(36) Mattar, S. M. *J. Phys. Chem. A* **2007**, *111*, 251.

(37) Kaprzac, S.; Reviakine, R.; Kaupp, M. *J. Phys. Chem. B* **2007**, *111*, 811.

(38) Adamo, C.; Cossi, M.; Barone, V. *THEOCHEM* **1999**, *493*, 145.

(39) Rega, N.; Cossi, M.; Barone, V. *J. Chem. Phys.* **1996**, *105*, 11060.

(40) Barone, V.; Cimino, P. *Chem. Phys. Lett.* **2008**, *454*, 139.

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

(42) Brancato, G.; Rega, N.; Barone, V. *Theor. Chem. Acc.* **2007**, *117*, 1001.

(43) Brancato, G.; Rega, N.; Barone, V. *J. Am. Chem. Soc.* **2007**, *129*, 15380.

(44) Hermosilla, L.; Calle, P.; Garcia de la Vega, J. M.; Sieiro, C. *J. Phys. Chem. A* **2005**, *109*, 114.

(45) Hermosilla, L.; Calle, P.; Garcia de la Vega, J. M.; Sieiro, C. *J. Phys. Chem. A* **2005**, *109*, 7626.

(46) Hermosilla, L.; Calle, P.; Garcia de la Vega, J. M.; Sieiro, C. *J. Phys. Chem. A* **2006**, *110*, 13600.

(47) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.

(48) Peterson, K. A.; Kendall, R. A.; Dunning, T. H. *J. Chem. Phys.* **1993**, *99*, 1930.

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

(50) Barone, V.; Cimino, P.; Pavone, M. In *Continuum Solvent Models in Chemical Physics: from Theory to Applications*; Mennucci, B., Cammi, R., Eds.; Wiley & Sons, Ltd.: 2007; Chapter *2.2*, pp 145–166.

(51) Neese, F. *J. Chem. Phys.* **2003**, *118*, 3939.

(52) Rustad, J. R.; Felmy, A. R.; Hay, B. P. *Geochim. Cosmochim. Acta* **1996**, *60*, 1553.

(53) Deyer, P. J.; Cummings, P. T. *J. Chem. Phys.* **2006**, *125*, 144519.

(54) Mason, M. G.; van Holle, W. G.; Robinson, D. W. *J. Chem. Phys.* **1971**, *54*, 3491.

(55) Werner, H. J.; Pavel, R. *J. Chem. Phys.* **1980**, *73*, 2319.

(56) Doerksen, R. J.; Thakkar, A. J.; Koga, T.; Hayashi, M. *THEOCHEM* **1999**, *488*, 217.

(57) Rakitin, A. R.; Yff, D.; Trapp, C. *J. Phys. Chem.* **2003**, *107*, 6281.

(58) Rega, N.; Cossi, M.; Barone, V. *J. Am. Chem. Soc.* **1997**, *119*, 12962.

(59) Rega, N.; Cossi, M.; Barone, V. *J. Am. Chem. Soc.* **1998**, *120*, 5723.

CT800034C

JCTC Journal of Chemical Theory and Computation

CHARMM Additive All-Atom Force Field for Acyclic Carbohydrates and Inositol

Ganesh Kamath, Olgun Guvench, and Alexander D. MacKerell, Jr.*

Department of Pharmaceutical Sciences, 20 Penn Street HSF II, University of Maryland, Baltimore, Maryland 21201

Received January 17, 2008

This paper was withdrawn on November 11, 2008.

10.1021/ct800019u CCC: $40.75
Published on Web 04/12/2008

JCTC Journal of Chemical Theory and Computation

Car–Parrinello Molecular Dynamics Simulations of $CaCl_2$ Aqueous Solutions

Teodora Todorova,[†,‡] Philippe H. Hünenberger,[§] and Jürg Hutter*,[†]

Physical Chemistry Institute, University of Zurich, CH-8057 Zurich, Switzerland, and Laboratory for Physical Chemistry, Swiss Federal Institute of Technology (ETH), CH-8093 Zurich, Switzerland

Received November 7, 2007

Abstract: Car–Parrinello molecular dynamics (CPMD) simulations are used to investigate the structural properties of 1 and 2 molal (m) $CaCl_2$ aqueous solutions and, in particular, the radial distribution functions, coordination numbers, and dipole moments of water molecules in the first solvation shell. According to these simulations, the first solvation shell of the Ca^{2+} ion consists of six water molecules, that are characterized by an increased averaged dipole moment compared to that of bulk water, and a first-shell Ca−O radial distribution function peak at 2.39 Å. The results are compared to those of CPMD simulations of Ca^{2+} (no counterions), and no significant differences are found. This indicates that the homogeneous neutralizing background charge density implicitly included in simulations of non-neutral systems appropriately mimics the presence of the counterions (at least in terms of reproducing the solvation structure properties and for the box sizes considered). Classical molecular dynamics (MD) simulations of aqueous Ca^{2+} using varying box sizes confirm this suggestion. The CPMD simulations at 2 m concentration also reveal additional possibilities for the structural arrangement of water molecules and chloride ions around Ca^{2+}. In particular, they support the stability of Ca^{2+}-Cl^- (contact) and Ca^{2+}-H_2O-Cl^- (solvent-separated) ion pairs. In addition, the solvent-separated cation pair is found to occur in a deprotonated Ca^{2+}-OH^--Ca^{2+} form. The existence of such a species has, to our knowledge, never been invoked previously to account for experimental data on $CaCl_2$ solutions.

1. Introduction

Although the Ca^{2+} aqua-ion is of great importance in biology and a key component of natural ground waters, many details of its solvation structure remain controversial.[1] Even for such fundamental properties as the average coordination number (CN) and the average Ca−O distance (r[Ca−O]) for the first solvation shell, the results widely depend on the (experimental or theoretical) method of investigation. Challenges to experimental approaches involve factors such as the low atomic number of the element (relatively weak scattering center) and the uncertainty in modeling the scattering data (solute–solvent correlations only accounting for a very limited fraction of the measured scattering intensities). In addition the r[Ca−O] distance of about 2.5 Å leads to a partial occlusion of the corresponding peak by the broad O−O peak of water in diffraction studies, rendering the experimental determination of the hydration structure difficult, in particular in the dilute regime. Based on X-ray diffraction (XRD) experiments on aqueous solutions of calcium halides or nitrate, CNs ranging between 5.9 and 8 have been reported[2–8] at concentrations ranging (in molal units) between 1 and 6 m (Table 1). Neutron diffraction (ND) experiments yielded values ranging between 5.5 and 10 in the concentration range between 1 and 6.4 m.[7,9–11] Extended X-ray absorption fine structure (EXAFS) measurements resulted in CNs between 6.8 and 8[12,13] at concentrations ranging from 0.12 to 6 m. Recently, Megyes et al.[7] reported the results of combined XRD and ND experiments on 2.5

[*] Corresponding author. E-mail: hutter@pci.uzh.ch.

[†] University of Zurich.

[‡] Current address: Paul-Scherrer-Institute (PSI), 5232 Villigen, Switzerland.

[§] Swiss Federal Institute of Technology (ETH).

10.1021/ct700302m CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/15/2008

Table 1. Experimental Results on Ca^{2+} Solvation Previously Reported in the Literature Based on X-ray Diffraction (XRD), Neutron Diffraction (ND), and Extended X-ray Absorption Fine Structure (EXAFS) Studies[a]

ref	method	salt	concn (m)	CN	r[Ca−O] (Å)
2	XRD	$CaCl_2$	1.1	6.9	2.39
3	XRD	$CaCl_2$	3.3	8	2.40
3	XRD	$CaCl_2$	5.2	8	2.40
4	XRD	$Ca(NO_3)_2$	3.6	7	2.44
4	XRD	$Ca(NO_3)_2$	6.0	7	2.45
5	XRD	$CaCl_2$	1.0	6	2.42
5	XRD	$CaCl_2$	2.0	6	2.42
5	XRD	$CaCl_2$	4.5	6	2.42
6	XRD	$CaCl_2$	2.0	6	2.46
6	XRD	$CaBr_2$	1.5	8	2.46
6	XRD	CaI_2	1.5	8	2.46
7	XRD+ND	$CaCl_2$	2.5	6.2(ND)	2.43–2.46(ND)
7	XRD+ND	$CaCl_2$	2.5	6.5(XRD)	2.43–2.46(XRD)
7	XRD+ND	$CaCl_2$	4.0	5.9(XRD)	2.43(XRD)
8	XRD	$CaCl_2$	1.0	8	2.45
9	ND	$CaCl_2$	4.5	5.5	2.41
10	ND	$CaCl_2$	1.0	10	2.46
10	ND	$CaCl_2$	2.8	7.2	2.39
10	ND	$CaCl_2$	4.5	6.4	2.40
11	ND	$CaCl_2$	6.4	6.95	2.41
11	ND	$CaCl_2$	4.0	7.3	2.40
12	EXAFS	$CaCl_2$	0.12	8	2.46
13	EXAFS	$CaCl_2$	0.2	6.8	2.43
13	EXAFS	$CaCl_2$	6.0	7.2	2.44

[a] The method, salt investigated, salt concentration (in molal units), and the resulting first-shell coordination number (CN) and average Ca−O distance (r[Ca−O]) are indicated.

Table 2. Theoretical Results on Ca^{2+} Solvation Previously Reported in the Literature from Monte Carlo (MC), Classical Molecular Dynamics (MD),[a] Quantum Mechanical/Molecular Mechanical (QM/MM) Simulations, Quantum Chemical Statistical Mechanical (QMSTAT) Calculations, and Car−Parrinello Molecular Dynamics (CPMD) Simulations[b]

ref	method	salt	concn (m)	CN	r[Ca−O] (Å)
2	MD	$CaCl_2$	1.1	9	2.39
6	MD-P(Åqvist)	Ca^{2+}		8	2.40
6	MD-P(Bounds)	Ca^{2+}		9–10	2.51
6	MD-P(Gromos)	Ca^{2+}		8	2.46
7	MD	$CaCl_2$	2.5	6.5	2.48
7	MD	$CaCl_2$	4.0	6.2	2.48
8	ab initio/MP2	Ca^{2+}	1	8	2.46
8	ab initio/MP2	Ca^{2+}	2.5	6.9	2.43
8	ab initio/MP2	Ca^{2+}	4	5.8	2.43
8	ab initio/MP2	Ca^{2+}	6	5.1	2.45
14	MD	$CaCl_2$	1.1	9–10	2.49
15, 16	MD-T	Ca^{2+}	∼1	7.1	2.50
15, 16	MD	Ca^{2+}	∼1	9.2	2.47
15, 16	QM/MM (ab initio)	Ca^{2+}	∼1	7.6	2.46
15, 16	QM/MM(DFT)	Ca^{2+}	∼1	8.1	2.51
17	MD	Ca^{2+}		8	2.50
18	MD	Ca^{2+}		9.3	2.54
19	QM/MM	Ca^{2+}	0.28	8.3	2.45
19	MD	Ca^{2+}	0.28	9.2	2.47
20	MD-P	Ca^{2+}		7.9	2.50
21	MD-P	Ca^{2+}		7.2–7.7	2.42–2.46
22	MC-P	Ca^{2+}		7	2.3
23	PCM/MD-P (PCM)	Ca^{2+}		8.6	2.50
27	QM/MM (ONIOM-XS)	Ca^{2+}		6	2.53
28	QMSTAT	Ca^{2+}		6.9	2.50
25	CPMD	Ca^{2+}		6	2.45
31	CPMD	Ca^{2+}		7–8	2.64
32	CPMD	Ca^{2+}		6.2	2.43
32	CPMD	Ca^{2+}		7	2.43
32	CPMD	Ca^{2+}		8	2.44
present study	CPMD	$CaCl_2$	1	6–7	2.39
present study	CPMD	$CaCl_2$	2	6	2.39
present study	CPMD	Ca^{2+}	2	6	2.39

[a] The "T", "P", or "PCM" added to the method indicates inclusion of three-body terms, explicit polarization, or polarizable continuum model implicit solvation. [b] The method, salt investigated, salt concentration (in molal units) and the resulting first-shell coordination number (CN) and average Ca−O distance (r[Ca−O]) are indicated.

and 4 m $CaCl_2$ solutions. At 2.5 m concentration they found consistent results of 6.5 ± 0.2 (XRD) and 6.2 ± 0.3 (ND) for the CN with a r[Ca−O] distance of 2.43–2.46 Å. At 4 m concentration the experiments suggested a CN of 5.9 ± 0.3 (XRD) with a r[Ca−O] distance of about 2.43 Å (XRD). Theoretical studies also contributed to the interpretation of experimental data and provided further (experimentally inaccessible) information on Ca^{2+} solvation (Table 2). The results also presented an important sensitivity to the applied methodology and parameters. Most empirical force field calculations relied on pairwise-additive interactions (no explicit electronic polarization). Molecular dynamics (MD) simulations with a modified central force–potential for water, together with ion–water and ion–ion pair potentials derived from ab initio calculations, suggested a first hydration shell consisting of either 9 or 10 water molecules and a r[Ca−O] distance of 2.49 Å for 1.1 m aqueous $CaCl_2$ solution.[14] In the higher concentration regime, CNs of 6.5 and 6.2 were suggested for 2.5 and 4.0 m $CaCl_2$, respectively,[7] together with a r[Ca−O] distance of 2.48 Å. Other simulations based on simple ab initio pair potentials gave CNs of 8–9.3 and r[Ca−O] distances of 2.39–2.54[2,15–19] in a similar concentration range. The use of empirical potentials including electron polarizability explicitly produced CNs between 7.2 and 10 and r[Ca−O] distances of 2.42–2.51 Å.[6,20,21] Monte Carlo (MC) simulations utilizing the MCHO potential form (including polarization) suggested values of 7 for the CN and 2.3 Å for the r[Ca−O].[22] The importance of electronic polarizability was investigated more directly in a comparison of classical simulations with pairwise additive versus explicitly polarizable ab initio derived ion–water potentials. The nonpolarizable model resulted in a CN of 9.2 and r[Ca−O] distance of 2.47,[19] while variants of the polarizable one gave CNs ranging between 7.2 and 8.6 with r[Ca−O] distances between 2.42 and 2.50 Å.[21,23] A good method to assess the importance of including higher-order terms (three-body terms or explicit polarizability) in the expansion of the interaction potential is to perform ab initio calculations of small ion–water clusters, revealing differences in binding energy per additional water molecule upon increasing the cluster size.[24,25] Such ab initio molecular orbital calculations, performed at the restricted Hartree–Fock (HF) and second-order Mϕller–Plesset perturbation (MP2) levels of theory and followed by natural energy decomposition analysis, emphasized the importance of polarization effects in the binding energies of $M^{2+}(H_2O)_n$ clusters.[26] Polarization has also been recognized through classical MD calculations using the polarizable continuum model (PCM) to be largely responsible for the nonclassical bent and pyramidal structures of the gas-phase di- and trihydrates and suggested to represent an important factor for determining the CN in solution.[23] Besides clusters in gas-phase, quantum-mechanical calculations also investigated solvation in the bulk. A comparison of (i) a classical simulation with a pairwise force field, (ii) a classical simulation with a force field including three-body terms, (iii) an ab initio quantum mechanical/molecular mechanical (QM/MM) simulation, and (iv) a density functional theory (DFT) QM/MM simulation reported CNs of 9.2, 7.1, 7.6, and 8.1, respectively,[15,16] with corresponding r[Ca−O] distances of 2.47, 2.50, 2.46, and 2.51 Å, respectively. QM/MM calculations with the ONIOM-XS (n-layered integrated molecular orbital and molecular mechanics extended to solvation) method led to a CN of 6 and an r[Ca−O] distance of 2.53 Å.[27] A CN of 6.9 together with an r[Ca−O] distance of 2.50 Å was also reported in a recent combined quantum chemical statistical mechanical (QMSTAT) calculation.[28] Finally, DFT based simulations using the Car–Parrinello molecular dynamics (CPMD) method[29,30] using different simulations protocols and density functionals (no counterions) showed variations in the water coordination number of Ca^{2+} from 6 to 8 and in the r[Ca−O] distance from 2.44 to 2.64 Å.[25,31,32]

An important point to keep in mind when performing classical or Car–Parrinello (CP) MD simulations to investigate the solvation properties of ions is that the finite size of the simulated systems and the approximate treatment of electrostatic interactions may have a significant impact on the simulation results.[33] In most cases, these simulations are performed under periodic boundary conditions (so as to eliminate surface effects) based on truly microscopic box volumes. When electrostatic interactions are handled as exactly periodic by application of lattice-sum methods (classical simulations) or plane-wave expansions (CP simulations), interactions between the reference box and its periodic images represent a significant (volume-dependent) perturbation compared to the ideal situation of a macroscopic solution at infinite dilution. Furthermore, in the case of non-neutral systems (e.g., single solvated ion without counterions), application of periodic electrostatics implicitly amounts to including a homogeneous neutralizing background charge density within the simulated box. The impact of periodicity induced artifacts on the solvation thermodynamics of ions is extremely large and cannot be neglected.[34–36] However, the associated structural perturbation (e.g., on pair distribution functions and CNs) may be more limited,[33,37] especially when the system is neutralized by the explicit inclusion of counterions. Nevertheless, the possible influence of artificially periodic electrostatics and neutralizing background charge on the results of previously reported CPMD simulations[25,31,32] of non-neutralized Ca^{2+} in water (which may still be important in view of the very small boxes considered in CP simulations compared to classical ones) has not been investigated systematically.

In the present study we expand on previously reported CPMD simulations.[25] The goal of this additional investigation is threefold: (i) comparing the results of a previous CPMD simulation[25] of 1 m Ca^{2+} in water (single Ca^{2+}, no counterions) with those of a new CPMD simulation of 1 m $CaCl_2$ (single Ca^{2+} and $2Cl^-$ counterions); (ii) evaluating via classical MD simulations the impact of the box size (via the homogeneous background charge density) on the simulated structural properties obtained from this previous CPMD simulation;[25] and (iii) investigating ion association properties in $CaCl_2$ solutions based on four new CPMD simulations of 2 m $CaCl_2$ ($2Ca^{2+}$ and $4Cl^-$) initiated from different starting configurations (free ions or Ca^{2+}-OH^--Ca^{2+}, Cl^--Ca^{2+}-Cl^- and Ca^{2+}-Cl^- species) along with one new CPMD simulation of 2 m Ca^{2+} ($2Ca^{2+}$; no counterions).

2. Computational Details

The CPMD[38] simulations were performed using the BLYP functional, i.e. with the exchange functional of Becke[39] and the correlation functional of Lee, Yang, and Parr,[40] using a Troullier-Martins norm-conserving pseudopotential for Ca^{2+} [41] and oxygen and hydrogen pseudopotentials as in ref 25. The mass of the hydrogen nucleus was set to that of the deuterium isotope. The valence electron wave function was expanded in plane waves with an energy cut off of 70 Ry. The fictitious electron mass was set to 600 au and the time step to 5 au (0.12 fs). The six simulations (see below) were carried out under periodic boundary conditions and involved 3 ps equilibration in the canonical ensemble, followed by 7.2 ps production in the microcanonical ensemble (the 1 m $CaCl_2$ production simulation was later extended to 19.2 ps). During the equilibration, the temperature was maintained at 320 K using a Nosé-Hoover thermostat.[42–45]

The systems consisted of a cubic box of 12.43 Å edge length, containing 58 water molecules (effective water density 1 g.cm^{-3}) and either 1 Ca^{2+} and 2 Cl^- ions (1 m $CaCl_2$) or $2Ca^{2+}$ and 4 Cl^- ions (2 m $CaCl_2$). A third system contained 62 water molecules, $2Ca^{2+}$ and no counterions (2 m Ca^{2+}). The resulting exact molalities and solution densities are 0.96 m and 1.00 g.cm^{-3} (1 m $CaCl_2$), 1.91 m and 1.09 g.cm^{-3} (2 m $CaCl_2$), and 1.79 m and 1.00 g.cm^{-3} (2 m Ca^{2+}), respectively. Note that the electrostatic interactions are treated as exactly periodic (i.e., they include contributions from the interaction between nuclei and electrons within the computational box as well as interactions between this box and all its periodic copies). In the case of a non-neutral system (previous study[25] of 1 m Ca^{2+} and present simulation of 2 m Ca^{2+}), this implicitly involves the inclusion of a homogeneous neutralizing background charge density within the computational box. When the system is neutral (present simulations of 1 and 2 m $CaCl_2$), there is no such background charge.

The initial configuration for the 1 m $CaCl_2$ system was taken from a classical MD simulation using the ion–water potential of Floris et al.[23] and the extended simple point charge SPC/E water model.[46] This initial configuration contained seven water molecules in the first solvation shell of the ion. The simulations of the 2 m $CaCl_2$ system were initiated from four different configurations illustrated in Figure 1. The initial configuration of System A consisted of two six-coordinated Ca^{2+} ions at a distance of 4 Å (free

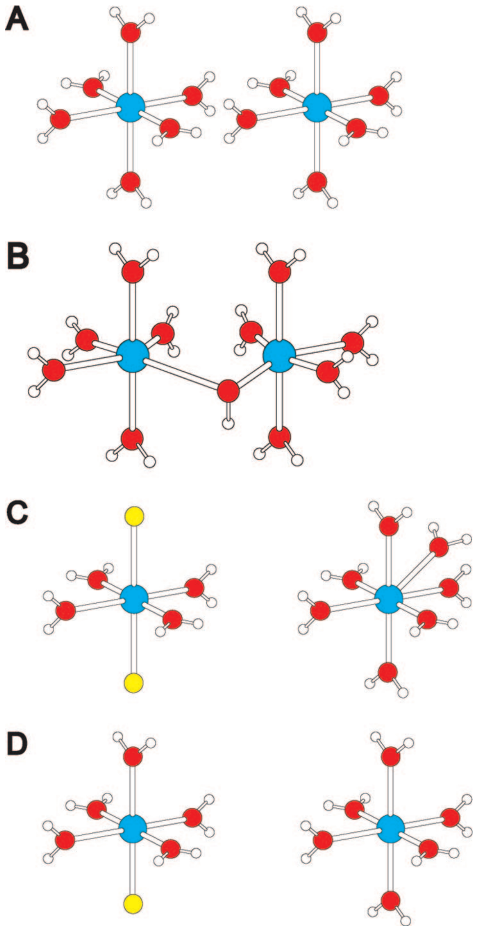


Figure 1. Initial configurations of the Car–Parrinello simulations (10.2 ps) of aqueous $CaCl_2$ ($2CaCl_2 + 58H_2O$; 2 m $CaCl_2$) corresponding to the simulated systems A–D (see Computational Details). A fifth system, System E, was also simulated in the absence of counterions ($2Ca^{2+} + 62H_2O$; 2 m Ca^{2+}). The corresponding initial configuration is identical to that of System A.

ions). The initial configuration of System B was generated from that of System A by slightly increasing the distance between the ions. This initial configuration consists of two six-coordinated Ca^{2+} ions at a distance of 4.5 Å, bridged by an OH^- group (species further noted Ca^{2+}-OH^--Ca^{2+}). The second hydrogen of the bridging water molecule formed an H_3O^+ ion in the bulk water. The initial configuration of System C was taken from a classical MD simulation using the ion–water potential of the GROMOS96 force field[47,48] and the SPC water model[46] Here, the Ca−Ca distance is 6.5 Å. While one of the Ca^{2+} ion is seven coordinated, the other forms axial Ca^{2+}-Cl^- bonds with two counterions, four water molecules being placed in equatorial positions (species further noted Cl^--Ca^{2+}-Cl^-). The initial configuration of System D was generated from that of System C by shortly increasing the distance between the Ca^{2+} and one Cl^- of the Cl^--Ca^{2+}-Cl^-species. This initial configuration consists of one six-coordinated ion (the seventh water having left the first coordination shell), the other ion forming one Ca^{2+}-Cl^- bond and presenting five water molecules in the first solvation shell (species further noted Ca^{2+}-Cl^-). Finally, a simulation was also performed for a System E containing $2Ca^{2+}$ ions (no Cl^-) and 62 water molecules. The initial structures of systems B−D were chosen in view of the role played by ion pairing (i.e., Ca^{2+}-Cl^-, Cl^--Ca^{2+}-Cl^-, and Ca^{2+}-H_2O-Cl^- species) in determining the thermodynamical properties of $CaCl_2$ solutions at high concentrations.[7,8,13,49] However, the Ca^{2+}-H_2O-Ca^{2+} species turned out to equilibrate to Ca^{2+}-OH^--Ca^{2+}.

The CPMD simulations were analyzed in terms of radial distribution functions (RDFs, $g(r)$), running coordination numbers (CNs, $n(r)$), time series of calcium−oxygen distances, and water dipole moment distribution in the first solvation shell. To examine the latter property, a localized molecular orbital analysis was performed by unitary transformation of Bloch orbitals to yield maximally localized Wannier functions.[50–52] Within the framework of the plane waves pseudopotential model, the maximally localized Wannier functions are analogous to the orbitals obtained by the Boys localization procedure commonly used in quantum chemistry.[53] The centroids of the Wannier functions (Wannier centers) were then used to define a dipole moment for individual water molecules in the simulations.

In addition to the CPMD simulations, a number of classical MD simulations were performed in order to evaluate the influence of modeling the counterion atmosphere implicitly by a homogeneous neutralizing background charge density in (CP or classical) simulations of Ca^{2+} under periodic boundary conditions (without explicit counterions). These simulations were performed at constant volume and temperature using the GROMOS96 program[47,48] together with the GROMOS 43a1 Ca^{2+} ion–solvent parameters and the SPC water model. The temperature was maintained close to 300 K by a Berendsen thermostat with a coupling time of 0.1 ps.[54] The systems involved one Ca^{2+} ion and a number of water molecules ranging from 8 to 150 in cubic boxes of edge lengths ranging from 6.48 to 16.59 Å (effective water density of about 1 $g.cm^{-3}$, omitting the ion). The electrostatic interactions were calculated using a lattice-sum method (P^3M,[55–57] with a sixth-order truncated-polynomial charge-shaping function of width 0.5 nm and a grid size of 32 × 32 × 32 points). This approach is the analog at the classical level of the periodic electrostatics used in the CPMD simulations. The cutoff radius for the van der Waals interactions was set to 10 Å. A multicell approach[58] was used to keep all parameters of the simulation constant when going to small simulation boxes (i.e., where the cutoff may exceed the half-box edge). The geometry of the water molecule was held rigid using SHAKE[47] with a relative geometric tolerance of 10^{-4}. A time step of 2 fs was used for integrating the equations of motion. All systems were equilibrated during 10 ps, followed by 500 ps production.

3. Results and Discussion

3.1. CPMD Simulations of a 1 m $CaCl_2$ Aqueous Solution. The Ca−O and Ca−H radial distribution functions (RDFs) as well as the running coordination numbers (CNs) obtained from the 7.2 ps CPMD simulation of a 1 m $CaCl_2$ solution at 320 K are displayed in Figure 2. The positions of the first maxima in the Ca−O and Ca−H RDFs are 2.39 Å and 3.03 Å, respectively, to be compared with correspond-

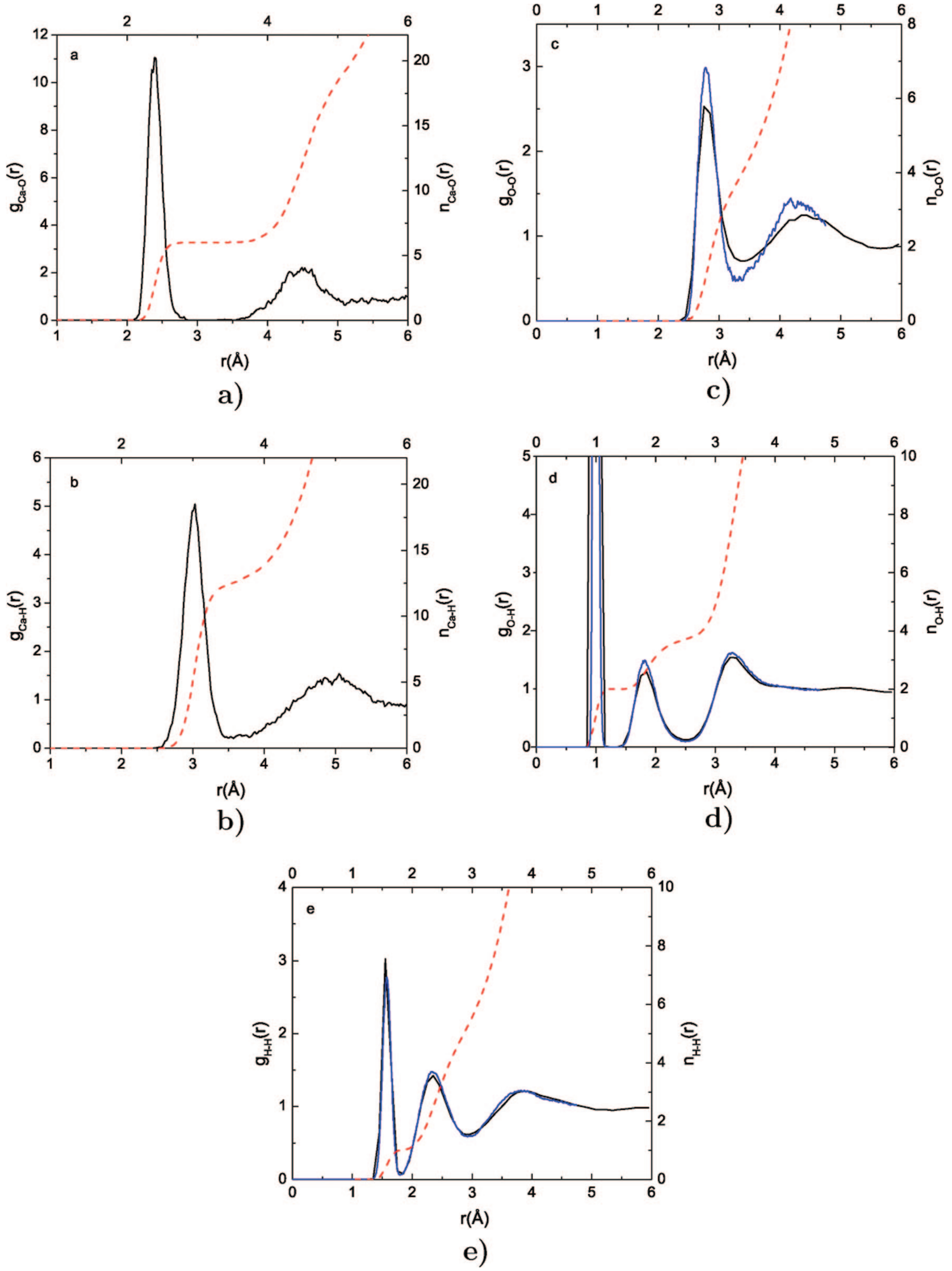


Figure 2. Radial distribution functions (RDFs; $g(r)$; black and blue lines) and running coordination numbers (CN, $n(r)$; dashed red lines) corresponding to calcium−water and water–water distances for the Car–Parrinello simulation (7.2 ps) of aqueous $CaCl_2$ (1 $CaCl_2$ + 58H_2O; 1 m $CaCl_2$; black and dashed red lines) and for Car–Parrinello simulations of pure water at the same level of theory (blue lines): a) Ca−O, b) Ca−H, c) O−O, d) O−H, e) H−H.

ing values of 2.45 Å and 2.98 Å for the previously reported CPMD simulation of 1 m Ca^{2+} without counterions.[25] For the Ca−O RDF the first and the second solvation shells are clearly separated by a region of nearly zero oxygen density, the two shells extending from about 2.1 to 2.8 Å and from about 3.5 to 5.1 Å, respectively. For the Ca−H RDF, the regions corresponding to the two solvation shells are broader, shifted to larger distances and partially overlapping. This difference between the two types of RDFs results from the preferential orientation of the water molecules (especially in the first shell) and the librational motions of the water molecules (especially in the second shell). Integrating the

Ca−O RDF over the first shell leads to an average first-shell CN of 6 (the corresponding integral over the Ca−H RDF is, as expected, about 12). Integrating the Ca−O RDF over the second shell leads to an average estimate of 13.2 water molecules. At 1 m $CaCl_2$ concentration the Cl^- counterions are found to systematically avoid the first and the second solvation shells of the Ca^{2+} ion (data not shown), so that their influence on the Ca^{2+} solvation structure is essentially negligible.

A comparison of the water–water O−O, O−H, and H−H RDFs with those obtained from CPMD simulation of pure water at the same level of theory[25] is also shown in Figure 2. The three functions evidence a loss of structure (particularly visible for the O−O RDF) in the regions corresponding to nearest and second nearest water molecules upon going from pure water to the $CaCl_2$ solution. The peaks corresponding to second nearest water molecules are also shifted to slightly larger distances. The first peaks in the O−O, O−H, and H−H RDFs are centered at 2.75, 0.95, and 1.55 Å, respectively (the former corresponding to intermolecular and the latter two to intramolecular distances).

The dipole moments of the six water molecules in the first coordination shell of Ca^{2+} were estimated based on the Wannier function approach as implemented in the CPMD code. This approach was previously applied[59] to calculate the dipole moment of the water molecule in the gas phase (1.86 D) and in the bulk (about 3 D). In aqueous $CaCl_2$ solutions, the dipole moment distribution for first-shell water molecules is shifted compared to that of pure water. It shows a maximum at 3.35 D, due to the electron polarization caused by the strong electric field of the ion. The same effect was also observed for Mg^{2+} and Be^{2+} ions in aqueous solution, with average first-shell water dipole moments of 3.3 and 3.1 D, respectively.[60,61]

The above results are very similar to those obtained in the previous investigation of a 1 m Ca^{2+} aqueous solution (no Cl^- counterions).[25] The counterions exert no noticeable effect on the first and second solvation shells of the Ca^{2+} ion (at the 1 m concentration considered here) and the first-shell CN remains six as found previously. This is also a first indication that the uniform neutralizing background charge density used in the CPMD simulation of the charged system[25] did not significantly affect the results (compared to the present explicit counterion treatment).

It has been suggested[21] that the low CNs and their sensitivity to the methodology employed in different CPMD simulations might be related to the fictitious electron mass used in these simulations.[25,31,32] We think that this is most likely not the case. The small mass of 600 au employed in this study produced the same results as in previous simulations (900 au fictitious electron mass in ref 25). In addition it has been convincingly shown[62,63] that no direct effect of the electron mass on static properties exists, if the masses employed stay within reasonable values and the simulations are correctly performed.

On the other hand, an important factor possibly affecting the CNs observed in different studies is the pseudopotential used to describe the core electrons of Ca^{2+}. Our tests show that it is important to include the $3p^6$ electronic state into the explicitly described valence electrons. A too soft Ca^{2+} ion results if these electrons are treated as part of the core. This has the effect that a larger water coordination shell can form and a CN of 7–8 is obtained.[31] Another important point is the choice of the exchange-correlation functional. Simulations utilizing the Perdew-Burke-Ernzerhof (PBE)[64] functional gave a CN of 6–7 for a flexible and 8 for a rigid model of water.[32] Recent calculations indicated that both PBE and BLYP exchange-correlation functionals produce too low water self-diffusion coefficients when compared to the experimental values.[65] In addition, it has been shown[66,67] that, at experimental density, both functionals lead to a very high pressure. Generalized gradient approximation density functionals, like BLYP and PBE, produce higher dipole moments when compared to dipole moments calculated with hybrid functionals. Furthermore, the isotropic polarizabilities are too large for most of these functionals. This is due to the well-known tendency of these density functionals to underestimate the HOMO–LUMO gap of molecular systems which, in turn, results in an overestimation of the molecular polarizability. Thus, the higher BLYP dipole moments and polarizabilities lead to lower CNs.

Another consequence of the overstructuring and slow diffusional dynamics of water when employing the BLYP representation is that only a few exchanges of molecules between the first and the second solvation shells are observed on the time scale of the present simulations. This feature is illustrated in Figure 3 (based on a trajectory extended to 19.2 ps). The initial configuration of the present simulation (pre-equilibrated using classical MD) presents seven water molecules in the first solvation shell of Ca^{2+}. The seventh water molecule leaves the shell shortly after the beginning of the simulation. The resulting six-coordinated Ca^{2+} structure remains stable until 13 ps and dominates the RDFs previously discussed (Figure 2; based on the initial 7.2 ps of the extended simulation). At this point, however, another water molecule enters the first solvation shell increasing again the CN to seven. Note that the seven-coordinated structures observed at the beginning and at the end of the simulation show more significant fluctuations in the first shell Ca−O distances (Figure 3) suggesting a destabilization of the solvation structure.

The results of the classical MD simulations, performed in order to investigate the box size dependence (via the background charge) of the structural properties in simulations of charged systems,[25] are reported in Table 3. Calculations with systems containing 8 to 150 water molecules and a single Ca^{2+} ion (no Cl^- counterion) have been performed at constant water density. The results show an effect on the coordination of Ca^{2+} for simulations with 50 water molecules or less, where the CN is observed to increase with the box size. This effect can be explained by the fact that the solvent molecules around the reference ion are perturbed by the periodic copies of the ion (undersolvation), the magnitude of the effect decreasing with increasing box size. An alternative (equivalent) interpretation is that the fraction of the homogeneous background charge that is inside the ionic volume (and thus reduces the field it exerts on the solvent) decreases upon increasing the box volume. The Ca−O

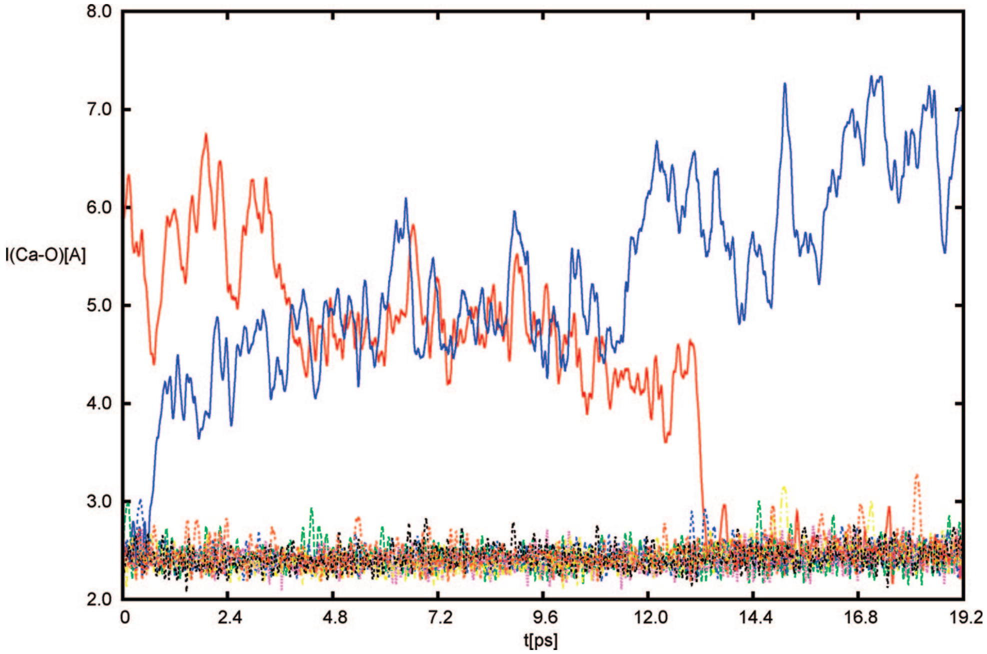


Figure 3. Time evolution of Ca−O distances along the Car–Parrinello simulation (19.2 ps; extended simulation) of aqueous $CaCl_2$ ($1CaCl_2 + 58H_2O$; 1 m $CaCl_2$). Only distances corresponding to the water molecules that belong to the first solvation shell at any time along the simulation are displayed.

Table 3. Coordination Number (CN) and First-Shell RDF Peak Position for the Ca−O Distance (*r*[Ca−O]) for Classical Simulations of Aqueous Ca^{2+} Solutions Involving 1 Ca^{2+} and the Indicated Number of Water Molecules[a]

H_2O	CN	*r*[Ca−O] (Å)
8	6.2	2.41
10	6.5	2.41
20	7.0	2.43
30	7.4	2.45
32	7.4	2.45
40	7.7	2.45
50	7.8	2.45
54	7.9	2.47
55	7.9	2.47
58	7.9	2.47
100	7.9	2.47
111	7.9	2.47
150	7.9	2.47

[a] At a constant effective water density of about 1 $g.cm^{-3}$, excluding the ion.

distance also increases along with the CN, due to the repulsion of the water molecules entering the first solvation shell. The same qualitative trend was found by Piquemal et al.[21] using a polarizable force field. In this study an increase of the CN from 7.2 to 7.7 was observed upon increasing the system size from 60 to 216 water molecules. However, this observation may also have other causes, the present results suggesting that finite-size effects on the solvation structure are essentially negligible above 50 water molecules.

3.2. CPMD Simulations of a 2 m $CaCl_2$ Aqueous Solution. The Ca−O and Ca−H RDFs as well as the running CNs obtained from the 7.2 ps CPMD simulations of a 2 m $CaCl_2$ (or Ca^{2+}) solution at 320 K are displayed in Figure 4 for the five different systems A−E. All systems remained close to the initial configuration during the entire simulation, and the final configurations of the runs were topologically identical to those displayed in Figure 1, justifying the calculation of average properties for the species considered (free ions, Ca^{2+}-OH^--Ca^{2+}, Cl^--Ca^{2+}-Cl^-, and Ca^{2+}-Cl^-) based on the different simulations. The positions of the first maxima in the Ca−O RDF range from 2.37 Å to 2.41 Å (to be compared with 2.39 Å for the simulation of 1 m $CaCl_2$, Figure 2, and 2.45 Å for the previous simulation[25] of 1 m Ca^{2+}). The first peaks corresponding to System A (2 m $CaCl_2$ solution) and System E (2 m Ca^{2+} solution; free ions) nearly exactly coincide, suggesting that the substitution of the chloride counterions by the uniform background charge density has little effect on first-shell solvation, at least when the two ions behave as independently solvated entities (i.e., at large enough distance). Note that System B has a slightly lower first peak and a second coordination shell shifted to smaller distances compared to System A. System C shows an even more pronounced decrease in the magnitude of the first peak along with a slightly altered curve shape for the second shell. Integrating the Ca−O RDFs up to the first minimum leads to average first-shell CNs ranging from 4 to 7 for the individual ions in the five different systems (Table 3). This number is about six for Systems A, B, and E. System C, representing the case of one seven-coordinated Ca^{2+} ion and another forming two Ca^{2+}-Cl^- bonds (d = 2.7–2.8 Å) and presenting four water molecules in equatorial positions, has the lowest RDF peak and an averaged CN of 5.5. The same averaged CN is also found for System D, but the RDF peak is higher and narrower, resulting from the smaller Ca−O distances compared to those of System C. The Ca−H radial distribution functions show the same trends (Figure 4, Table 3): almost no difference between Systems A, B, and E (CN of about 12 for the H atoms) and a reduced coordination for Systems C and D, due to the missing water molecules in the first solvation shell. The averaged dipole

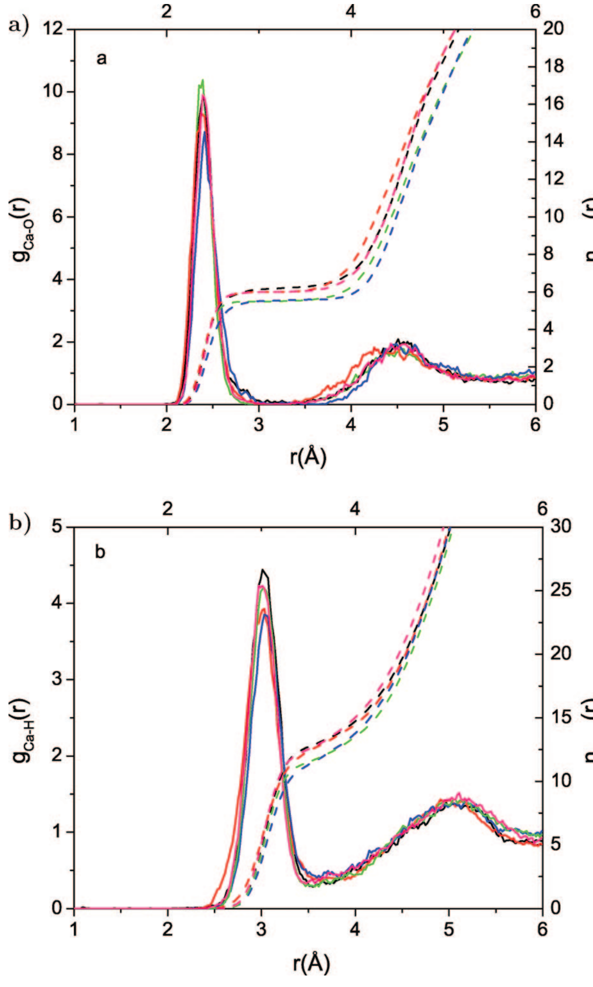


Figure 4. Radial distribution functions (RDFs; *g*(*r*), solid lines) and running coordination numbers (CNs, *n*(*r*), dashed lines) corresponding to Car–Parrinello simulations (7.2 ps) of aqueous $CaCl_2$ (Systems A−D; $2CaCl_2 + 58H_2O$; 2 m $CaCl_2$) or Ca^{2+} (System E; $2Ca^{2+} + 62H_2O$): a) Ca−O and b) Ca−H: System A (black), System B (red), System C (blue), System D (green), System E (pink), see Computational Details.

moment calculated for the water molecule in the first solvation shell are also reported in Table 3. The results show two trends: first, the higher the CN, the lower the dipole moment; second, the higher the concentration, the lower the dipole moment. A probable explanation is that at higher concentrations the Ca^{2+} hydration structure is weakened by the perturbation caused by the neighboring Ca^{2+} and Cl^- ions and the water polarization is reduced, resulting in lower dipole moments. A comparison of Systems A, C, and D with System E shows that even at 2 m concentration there is no significant influence of the chloride ions on the dipole moment distributions.

The structures of all systems (Figure 1) were preserved during the CPMD simulations. In System B, there was no water exchange between the first and the second shells, and the Ca^{2+}-OH^- distance remained at all times close to 2.39 Å. The coordination structure of the first ion in System C, involving two axial Ca^{2+}-Cl^- bonds and four water molecules in equatorial positions was also stable. However, the second ion was seven-coordinated, and the Ca−O distance oscillations were found to be larger (in analogy with Figure 3, although no exchange of water involving the first solvation shell was observed in this case).

Measurements on aqueous solutions of $MgCl_2$ and $CaCl_2$ showed an anomalous behavior of the osmotic coefficient for $CaCl_2$ (but not for $MgCl_2$) at about 5 m concentration.[49] This observation was suggested to result from the formation of Ca^{2+}-Cl^- species (contact ion pairs) at high concentrations. The formation of contact ion pairs has indeed been observed in high temperature aqueous solutions,[68–70] probably due to the reduced dielectric permittivity of water (leading to enhanced ion association). However, EXAFS studies suggested that Ca^{2+}-OH_2-Cl^- species (solvent separated ion pairs) rather than contact ion pairs were responsible for the anomaly of the osmotic coefficient at 5 m concentration.[13] Different molal concentrations appear to result in the dominance of different species: solvent separated pairs at 6 m concentration and contact pairs at 9.2 m concentration.[13] Average Ca−O distances were estimated to be 2.44 Å, and average Ca−H distances to be 2.97 Å.[13] Independent ND experiments[11] confirmed many of the findings of this EXAFS study.[13] Specifically: minimal changes in the nearest-neighbor Ca−O correlation, a mean CN of about 7 for the first hydration shell, and the absence of significant Ca^{2+}-Cl^- contact ion pairing even at concentrations as high as 6.4 m.[11] Chloride ions were proposed to be in the second coordination shell (Ca^{2+} to Cl^- distances of 4.6–5.6 Å) with the dominance of Ca^{2+}-OH_2-Cl^- solvent-separated pairs at 4 and 6.4 m. XRD experiments also suggested the dominance of these species at 1 m and the simultaneous presence of both Ca^{2+}-OH_2-Cl^- (solvent separated) and Ca^{2+}-Cl^- (contact) ion pairs at 4 and 6 m.[8] XRD experiments reported first-shell Ca−O distances in the range 2.43–2.46 Å, Ca^{2+}-Cl^- distances of about 2.75 Å and a Cl^--O distance of about 3.25 Å at 4–6 m.[8] Recently, Megyes et al.[7] performed XRD, ND, and MD studies and discussed the relative abilities of these methods to detect the ion pair formation. They point out that XRD could detect ion pairs in the case of concentrated solutions (more than 4 m), while ND required very careful isotope substitution in order to be able to detect ion pairs in solution. The MD results were in general accordance with the XRD experimental findings.[7]

The present study suggests that the Ca^{2+}-Cl^- (contact) or Ca^{2+}-H_2 O−Cl^- (solvent separated) ion pairs may be present even at lower concentrations (2 m), at least as transient species stable on the 10 ps time scale probed by the present simulations. However, their relative stabilities (i.e., equilibrium concentrations) cannot be evaluated based on such short simulations. In addition, a particularly interesting result is the observation of a Ca^{2+}-OH $^-$-Ca^{2+} solvent-separated species, instead of the Ca^{2+}-H_2 O-Ca^{2+} species. To the best of our knowledge, XRD, EXAFS, and ND experiments show no evidence for such a hydroxyl-separated ion pair. To rationalize System B, we looked at the thermodynamics of formation, deprotonation, and charge reduction of the hydrates of doubly charged ions. One of the methods to produce doubly charged metal ion hydrates is the clustering method. This method is based on a charge reduction, which can be viewed as a proton transfer reaction, i.e. $M(H_2O)^{2+} + H_2O \rightarrow M(OH)^{+} + H_3O^{+}$ in the gas phase (M: divalent metal ion). Density functional theory cluster calculations have shown that in the case of Ca^{2+}, the charge reduction reaction costs only 19.2 kcal/mol, while the competing reaction of water loss would require 48.1 kcal/mol.[71] Moreover, it has been proposed that water, with a proton affinity of 165 kcal/mol, is able to deprotonate $Ca(H_2O)^{2+}$. Because of the large release of kinetic energy associated with the Coulombic repulsion between the two positively charged ions produced by the charge reduction reaction, only bases whose proton affinity is significantly larger than the proton affinity of $CaOH^{+}$ (108 kcal/mol) will be able to induce a proton transfer.[72] A similar situation arises for the Al^{3+} ion where the kinetics of proton and water exchange in aqueous Al^{3+} support five-coordinated $Al(H_2O)_4OH^{2+}$ ions as the predominant form of $AlOH(aq)^{2+}$ under ambient conditions.[73]

Table 4. First-Shell Coordination Number (CN), First-Shell RDF Peak Position for the Ca–O and Ca–H Distances (*r*[Ca–O] and *r*[Ca–H]), and Average Dipole Moments of First-Shell Water Molecules for Car–Parrinello Simulations of Different Aqueous Ca^{2+} or $CaCl_2$ Solutions

system	atom	CN	*r*[Ca–O] (Å)	*r*[Ca–H] (Å)	dipole (D)
$Ca + 54H_2O$	Ca_1	6	2.45	2.98	3.40
$CaCl_2 + 58H_2O$	Ca_1	6(7)	2.39	3.03	3.35
A	Ca_1	6	2.39	3.01	3.25
A	Ca_2	6	2.39	3.01	3.25
B	Ca_1-Ca_2	12	2.37	3.03	3.15
C	Ca_1	7	2.41	3.03	3.15
C	Ca_2	4	2.41	3.03	3.25
D	Ca_1	6	2.39	3.03	3.20
D	Ca_2	5	2.39	3.03	3.25
E	Ca_1	6	2.39	2.97	3.25
E	Ca_2	6	2.39	2.97	3.25

4. Conclusions

In the present study, six CPMD simulations of aqueous $CaCl_2$ and Ca^{2+} solutions are reported.

The results of the simulation of a 1 m $CaCl_2$ solution ($1CaCl_2 + 58H_2O$) are consistent with those of a previous calculation on a 1 m Ca^{2+} solution ($1Ca^{2+} + 54H_2O$) using the same methodology.[25] Classical MD simulations of a single Ca^{2+} ion in boxes of increasing sizes confirm that beyond about 50 water molecules included in the computational box, the calculated structural properties (RDF peaks and CN) become essentially independent of the box size. The previous and present CPMD simulations support a first shell RDF peak for the Ca–O distance at 2.39 Å and a CN of 6 for the first solvation shell at a concentration of 1 m. However, the time frame of the simulations (about 10 ps) does not allow for a statistical sampling of water exchange between first and second solvation shells, and it is likely that the converged averaged CN using the present methodology would be slightly larger than 6. The second solvation shell of Ca^{2+} ranges from about 3.5 to 5.1 Å, and encompasses about 13.2 water molecules on average. Localized orbital analysis based on Wannier functions calculations shows a polarization of water molecules in the first solvation shell, which results in an increased dipole moment of 3.3 D compared to that of the bulk water (about 3 D). The tests performed in the present study thus indicate that there are no artifacts from simulation protocols using uniform background charges and artificial periodicity at the experimental concentration of 1 m. In former studies only the stability of the results toward infinite dilution was considered. The present results achieved with more strict parameters used in the Car–Parrinello simulations give more credit to previous simulations.[25,32]

Simulations of a 2 m $CaCl_2$ solution ($2CaCl_2 + 58H_2O$) or of a 2 m Ca^{2+} solution ($2Ca^{2+} + 62H_2O$) starting from free (i.e., widely separated) ions also produced very similar results (among each other and compared to the 1 m case). Recent combined XRD and ND experiments[7] suggest CN of 6.5 ± 0.2 and 6.2 ± 0.3 together with an average Ca–O distance between 2.43 and 2.46 Å for 2.5 m solutions, respectively. The slightly too low values for these quantities in our simulations are most likely due to shortcomings in the functional (BLYP) employed and not due to simulation protocol parameters or system size dependencies, as was recently[21] suggested. Another source of discrepancies might be that simulations performed at the experimental density result in an effectively overpressurized sample when using the BLYP functional. Finally, the simulations at 2 m $CaCl_2$ solutions ($2CaCl_2 + 58H_2O$) starting from four different configurations support the stability of Ca^{2+}-Cl^{-} (contact) and Ca^{2+}-OH_2-Cl^{-} (solvent-separated) ion pairs on the 10 ps time scale. This result is in agreement with recent XRD[8] and EXAFS[13] experiments. The solvent-separated cation pair was found to be present in the simulations in its deprotonated form Ca^{2+}-OH^{-}-Ca^{2+}. To the best of our knowledge, no such specie has been reported or suggested based on experiment. Although the above species were found to be stable on the 10 ps time scale, it is impossible to infer their relative stability (i.e., their equilibrium concentrations) from the present simulations. However, the simulations suggest that at higher concentrations, a wider variety of species may exist, and the study of these complex solutions by computational means makes the explicit inclusion of electronic structure necessary. On the other hand, is it also clear from our simulations that the scope of such first principles simulations is limited due to the restricted time frame currently available.

Acknowledgment. T.T. wants to thank Dr. A. P. Seitsonen and Dr. M. Iannuzzi for fruitful discussions. Calculations were performed on the Matterhorn Cluster at the University of Zürich.

References

(1) da Silva, J. F.; Williams, R. *The Biological Chemistry of the Elements*; Clarendon Press: Oxford, 1991; pp 278–314.

(2) Probst, M.; Radnai, T.; Heinzinger, K.; Bopp, P.; Rode, B. J. *Phys. Chem.* **1985**, *89*, 753–759.

(3) Albright, J. J. *Chem. Phys.* **1972**, *1972*, 3783.

(4) Smirnov, P.; Yamagami, M.; Wakita, H.; Yamaguchi, T. *J. Mol. Liq.* **1997**, *73–4*, 305–316.

(5) Licheri, G.; Piccaluga, G.; Pinna, G. J. *Chem. Phys.* **1976**, *64*, 2437–2441.

(6) Jalilehvand, F.; Spangberg, D.; Lindqvist-Reis, P.; Hermansson, K.; Persson, I.; Sandstrom, M. *J. Am. Chem. Soc.* **2001**, *123*, 431–441.

(7) Megyes, T.; Bakó, I.; Bálint, S.; Grósz, T.; Radnai, T. J. *Mol. Liq.* **2006**, *129*, 63–74.

(8) Megyes, T.; Grósz, T.; Radnai, T.; Bakó, I.; Pálinkás, G. J. *Phys. Chem. A* **2004**, *108*, 7261–7271.

(9) Cummings, S.; Enderby, J.; Howe, R. J. *Phys. C: Solid State Phys.* **1980**, *13*, 1–8.

(10) Hewish, N.; Neilson, G.; Enderby, J. *Nature* **1982**, *297*, 138–139.

(11) Badyal, Y.; Barnes, A.; Cuello, G.; Simonson, J. J. *Phys. Chem. A* **2004**, *108*, 11819–11827.

(12) Spangberg, D.; Hermansson, K.; Lindqvist-Reis, P.; Jalilehvand, F.; Sandstrom, M.; Persson, I. J. *Phys. Chem. B* **2000**, *104*, 10467–10472.

(13) Fulton, J.; Heald, S.; Badyal, Y.; Simonson, J. J. *Phys. Chem. A* **2003**, *107*, 4688–4696.

(14) Pálinkás, G. *Chem. Phys. Lett.* **1986**, *126*, 251–254.

(15) Schwenk, C.; Loeffler, H.; Rode, B. J. *Chem. Phys.* **2001**, *115*, 10808–10813.

(16) Schwenk, C.; Rode, B. *Pure Appl. Chem.* **2004**, *76*, 37–47.

(17) Obst, S.; Bradaczek, H. J. *Phys. Chem.* **1996**, *100*, 15677–15687.

(18) Periole, X.; Allouche, D.; Daudey, J.; Sanejouand, Y. J. *Phys. Chem. B* **1997**, *101*, 5018–5025.

(19) Tongraar, A.; Liedl, K.; Rode, B. J. *Phys. Chem. A* **1997**, *101*, 6299–6309.

(20) Kalko, S.; Sesé, G.; Padró, J. J. *Chem. Phys.* **1996**, *104*, 9578–9585.

(21) Piquemal, J.; Perera, L.; Cisneros, G.; Ren, R.; Pedersen, L.; Darden, T. J. *Chem. Phys.* **2006**, *125*, 054511.

(22) Bernal-Uruchurtu, M.; Ortega-Blake, I. J. *Chem. Phys.* **1995**, *103*, 1588–1598.

(23) Floris, F.; Persico, M.; Tani, A.; Tomasi, J. *Chem. Phys. Lett.* **1994**, *227*, 126–132.

(24) Pavlov, M.; Siegbahn, P.; Sandstrom, M. J. *Phys. Chem. A* **1998**, *102*, 219–228.

(25) Bakó, I.; Hutter, J.; Pálinkás, G. J. *Chem. Phys.* **2002**, *117*, 9838–9843.

(26) Glendening, E.; Feller, D. J. *Phys. Chem.* **1996**, *100*, 4790–4797.

(27) Kerdcharoen, T.; Morokuma, K. J. *Chem. Phys.* **2003**, *118*, 8856–8862.

(28) Tofteberg, T.; Ohrn, A.; Karlstrom, G. *Chem. Phys. Lett.* **2006**, *429*, 436–439.

(29) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(30) Marx, D.; Hutter, J. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; FZ Jülich: Germany, 2000; Vol. 1 of NIC Series, pp 329–477.

(31) Naor, M. M.; Nostrand, K. V.; Dellago, C. *Chem. Phys. Lett.* **2003**, *369*, 159–164.

(32) Lightstone, F. C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G. *ChemPhysChem* **2005**, *6*, 1745–1749.

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

(34) Kastenholz, M.; Hünenberger, P. H. J. *Phys. Chem. B* **2004**, *108*, 774–788.

(35) Kastenholz, M.; Hünenberger, P. H. J. *Chem. Phys.* **2006**, *124,* 124106.

(36) Kastenholz, M.; Hünenberger, P. H. J. *Chem. Phys.* **2006**, *124*, 224501.

(37) Peter, C.; van Gunsteren, W. F.; Hünenberger, P. H. J. *Chem. Phys.* **2002**, *116*, 7434–7451.

(38) *CPMD V3.11*; copyright IBM Corp 1990–2007, copyright MPI für Festkörperforschung Stuttgart 1997–2001.

(39) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

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

(41) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.

(42) Nosé, S. J. *Chem. Phys.* **1984**, *81*, 511–519.

(43) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(44) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(45) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. J. *Chem. Phys.* **1992**, *97*, 2635–2643.

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

(47) van Gunsteren, W.; Billeter, S.; Eising, A.; Hünenberger, P.; Krüger, P.; Mark, A.; Scott, W.; Tironi, I. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Hochschulverlag an der ETH Zurich/Biomos: Zurich/Groningen, 1996.

(48) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Billeter, A. E. M. S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. J. *Phys. Chem. A* **1999**, *103*, 3596–3607.

(49) Phutela, R.; Pitzer, K. J. *Solut. Chem.* **1983**, *12*, 201–207.

(50) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12865.

(51) Berghold, G.; Mundy, C. J.; Romero, A. H.; Hutter, J.; Parrinello, M. *Phys. Rev. B* **2000**, *61*, 10040–10048.

(52) Kirchner, B.; Hutter, J. J. *Chem. Phys.* **2004**, *121*, 5133–5142.

(53) Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 296–299.

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

(55) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles.* Institute of Physics Publishing: Bristol, 1981; pp 267–304.

(56) Hünenberger, P. H. Lattice-sum methods for computing electrostatic interactions in molecular simulations. In *Simulation and theory of electrostatic interactions in solution: Computational chemistry, biophysics, and aqueous solutions*. Pratt, L. R., Hummerand, G., Eds.; American Institute of Physics: New York, U.S.A., 1999; pp 17–83.

(57) Hünenberger, P. H. J. *Chem. Phys.* **2000**, *113*, 10464–10476.

(58) de Vries, A. H.; Chandrasekhar, I.; van Gunsteren, W. F.; Hünenberger, P. H. J. *Phys. Chem. B* **2005**, *109*, 11643–11652.

(59) Silvestrelli, P. L.; Parrinello, M. J. *Chem. Phys.* **1999**, *111*, 3572–3580.

(60) Lightstone, F. C.; Schwegler, E.; Hood, R. Q.; Gygi, F.; Galli, G. *Chem. Phys. Lett.* **2001**, *343*, 549–555.

(61) Marx, D.; Sprik, M.; Parrinello, M. *Chem. Phys. Lett.* **1997**, *273*, 360–366.

(62) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108*, 12990–12998.

(63) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I. *J. Chem. Theory Comput.* **2006**, *2*, 1274–1281.

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

(65) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I.-F. W.; Mundy, C. J. J. *Phys. Chem. B* **2006**, *110*, 3685–3691.

(66) McGrath, M. J.; Siepmann, J. I.; Kuo, I.-F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. *J. Phys. Chem. A* **2006**, *110*, 640–646.

(67) McGrath, M. J.; Siepmann, J. I.; Kuo, I.-F. W.; Mundy, C. J. *Mol. Phys.* **2006**, *104*, 3619–3626.

(68) Hoffmann, M.; Darab, J.; Palmer, B.; Fulton, J. J. *Phys. Chem. A* **1999**, *103*, 8471–8482.

(69) de Bakker, P. I. W.; Hünenberger, P. H.; McCammon, J. A. J. *Mol. Biol.* **1999**, *285*, 1811–1830.

(70) Peric, L.; Pereira, C. S.; Hünenberger, P. H. *Mol. Simul.* **2007**, submitted for publication.

(71) Peschke, M.; Blades, A.; Kebarle, P. *Int. J. Mass Spectrom.* **1999**, *187*, 685–699.

(72) Gronert, S. J. *Am. Chem. Soc.* **1996**, *118*, 3525–3526.

(73) Swaddle, T. W.; Rosenqvist, J.; Yu, P.; Bylaska, E.; Phillips, B. L.; Casey, W. H. *Science* **2005**, *308*, 1450–1453.

CT700302M

JCTC Journal of Chemical Theory and Computation

Electrostatically Embedded Multiconfiguration Molecular Mechanics Based on the Combined Density Functional and Molecular Mechanical Method

Masahiro Higashi and Donald G. Truhlar*

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

Received January 4, 2008

Abstract: We present a new method for generating global or semiglobal potential energy surfaces in the presence of an electrostatic potential; the new method can be used to model chemical reactions in solution or in an enzyme, nanocavity, or other chemical environment. The method extends the multiconfiguration molecular mechanics method so that the energy depends on the electrostatic potential at each atomic center. The charge distribution of the system can also be calculated. We illustrate the method by applying it to the symmetric bimolecular reaction $Cl^- + CH_3Cl' \rightarrow ClCH_3 + Cl'^-$ in aqueous solution, where the potential energy information is obtained by the combined density functional and molecular mechanical method, that is, by the combined quantum mechanical and molecular mechanical method (QM/MM) with the QM level being density functional theory. It is found that we can describe a semiglobal potential energy surface in aqueous solution with electronic structure information obtained entirely in the gas phase, including the linear and quadratic responses to variations in the electrostatic potential distribution. The semiglobal potential energy surface calculated by the present method is in good agreement with that calculated directly without any fitting.

1. Introduction

Combined quantum mechanical and molecular mechanical (QM/MM) methods have provided powerful means for studying chemical reactions in solution, enzymes, and solids.[1–28] In these approaches, the solute molecule or the reaction center involved in the formation and breaking of chemical bonds is described quantum mechanically, while the surroundings (e.g., solvent, solid surface, or protein environment) are treated by using a MM force field. When the system contains a large number of atoms, a statistical sampling method such as molecular dynamics (MD) or Monte Carlo simulation is required.

However, the high computational cost of *ab initio* or density functional QM calculations prevents carrying out QM/MM MD simulations with reliable accuracy and adequate sampling. To overcome this difficulty, many more approximate methods have been developed, but we can mainly classify them into three types. In the first type of method, a reaction path connecting the reactant and product is first determined in limited dimensionality, for example, in the gas phase or with nonquantal degrees of freedom (corresponding to spectator atoms or a secondary zone) excluded (in which case, the method is called QM-FE) or frozen (in which case it is called QM/MM-FE). Then, the free energy profile is obtained by free energy perturbation calculations along the path with the QM coordinates and electron density fixed.[29–36] These methods assume that the dynamics of the QM and MM subsystems are independent of each other and that the QM subsystem needs to be treated only in the quadratic region around the single uncoupled path.[37] Although several efficient algorithms for tracing the reaction paths have been developed,[30,33–35] this approach sometimes has a difficulty that the reaction path is trapped at one of the local minima of the potential energy surface (PES) and not smoothly connected from the reactant to product because there are many local minima on the MM PES.[32] Any single reaction path can deviate significantly from paths that make an appreciable contribution in a properly sampled thermal ensemble,[38–48] even if the single

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

10.1021/ct800004y CCC: $40.75
Published on Web 03/26/2008

path is the minimum-free-energy path (MFEP) on the potential of mean force (PMF) for a large subset of the degrees of freedom. (The PMF is an averaged energy surface, in particular, a free energy surface (FES). The true dynamics involve an average over paths, not the optimized path on an average surface, and even if the subset of the degrees of freedom included in the potential of mean force were large enough, this potential provides the full information needed to describe the dynamics only if classical transition state theory applies with a transmission coefficient of unity.) In addition, since the QM coordinates and charge distribution are fixed during MD simulations of this type, a significant part of the coupling between the QM and MM regions is ignored.

In the second type of calculation, high-level electronic structure methods combined with dielectric continuum models[49–54] or integral equation theories of solvation (such as the reference interaction site model self-consistent field (RISM-SCF) method[55–58]) are used to calculate the free energy surfaces of chemical reactions in solution. Although these methods do not need to sample the solvent degrees of freedom, they cannot easily be applied to reactions with inhomogeneous environments such as proteins, and furthermore they again yield only a preaveraged surface. (For many purposes, it is an advantage to directly calculate the FES, and it facilitates the calculation of equilibrium solvation paths[59,60] (ESPs), also called MFEPs,[61] and transition state theory rate constants,[53] but the PES required for full real-time dynamics can only be obtained from the FES by making further approximations.[62,63] Note that an ESP is a special case of a MFEP in which the primary coordinates on which the FES depends correspond to the coordinates of a solute or a microsolvated solute, and the secondary subsystem that is averaged corresponds to the solvent or the rest of the solvent.)

In the third type of calculation (SE-MO/MM), one uses proper free energy sampling of unaveraged motions, but due to cost, one uses semiempirical molecular orbital (SE-MO) methods such as Austin model 1 (AM1),[64] parametrized model 3 (PM3),[65] or self-consistent-charge density-functional tight binding (SCC-DFTB)[66,67] instead of high-level methods in the QM electronic structure calculation. Semiempirical methods require a much lower computational cost than *ab initio* or density functional methods, and direct SE-MO/MM dynamics simulations are feasible, so dynamical properties such as transmission coefficients can be calculated straightforwardly.[16,47] However, it is well-known that SE-MO is less reliable than *ab initio* wave function theory and density functional theory.

Lu and Yang[37] re-examined the QM/MM-FE method[30,34] and summarized its chief approximations as (i) assuming that the dynamics of the QM and MM subsystems are independent of each other and (ii) assuming that the QM system is confined to the quadratic region around the single uncoupled path. They then proposed a new method, called QM/MM-RPP where the PES and its electron density response properties are expanded to second order along a reaction path.[37] The expanded potential and response properties provide what may be called a reaction path potential (RPP), which is a concept widely used in gas-phase dynamics.[68–74] Yang and co-workers[61] subsequently extended the theory to optimize the reaction path on a FES; they call the resulting theory the QM/MM minimum free-energy path (QM/MM-MFEP) method. This method can treat the dynamical coupling between the QM and MM regions with QM/MM methods employing high-level QM in the vicinity of the MFEP. However, a second-order expansion is valid only near the origin of the expansion, and many expansion points are required to calculate a global FES. A global PES or global FES is needed to compute a broad distribution of reaction paths such as, for example, those often involved in large-curvature tunneling, which can make a significant contribution to the rate of hydrogen transfer reactions such as proton transfer reactions. For this purpose, and because even for the small-curvature tunneling case the ensemble of reaction paths can be broad,[47] it is desirable to develop a method to describe the global PES with a minimum of high-level QM input. This is the objective of the present study.

The multiconfiguration molecular mechanics (MCMM) method will be the starting point for the present development. MCMM has been successful in describing semiglobal potential energy surfaces of gas-phase reactions and calculating the reaction rates with multidimensional tunneling contributions.[75–82] In the MCMM method, which is compared elsewhere[83,84] (with more than 30 references) to related approaches, the Born–Oppenheimer potential energy at geometry **q** is represented as the lowest eigenvalue of the 2 × 2 diabatic Hamiltonian matrix:

$$\mathbf{V}^{\mathrm{MCMM}}(\mathbf{q}) = \begin{pmatrix} V_{11}(\mathbf{q}) & V_{12}(\mathbf{q}) \\ V_{12}(\mathbf{q}) & V_{22}(\mathbf{q}) \end{pmatrix} \quad (1)$$

where the diagonal elements, V_{11} and V_{22}, are MM energy functions that describe reactants and products, respectively. The off-diagonal element, V_{12}, and its derivatives are determined to reproduce high-level electronic structure calculations of the energy, gradient, and Hessian at some reference points called Shepard points, and modified Shepard interpolation[85,86] is used to interpolate the PES between the trust regions of the resulting set of second-order Taylor series. In the case of reactions with more than one possible product, MCMM would need to be extended, for example, to use a 3 × 3 matrix. The computational cost of using MCMM is much lower than that of using high-level electronic structure calculations directly.

In the present paper, we propose a method called electrostatically embedded multiconfiguration molecular mechanics (EE-MCMM). The new method is based on QM/MM methodology, and it extends the original MCMM by adding the electrostatic potential on each QM atom from the MM regions to V^{MCMM}. Taylor expansions are carried out with respect to both the nuclear coordinates and the electrostatic potentials at the nuclei; the coefficients of the Taylor series are determined such that they reproduce high-level electronic structure calculations at Shepard points. The collection of the values of the external electrostatic potential at the locations of the QM nuclei will be called the electrostatic potential distribution. The EE-MCMM allows us to calculate the PES in the presence of an external electrostatic potential. The Taylor series can represent the electrostatic potential due to the MM subsystem, and thus EE-MCMM can describe semiglobal PESs with moderate computational cost. Because the method is efficient, we can use DF/MM, that is, QM/MM with the QM level being density functional theory.

We illustrate the new method by application to the symmetric bimolecular reaction $C1^- + CH_3Cl' \rightarrow ClCH_3 + Cl'^-$ in aqueous solution, a reaction that has been investigated with various theoretical methods.[29,87–108] We first create a semiglobal PES in the gas phase by MCMM. The PES generated by MCMM is compared to that calculated directly without any fitting in a wide swath from the reactant through the saddle point (SP) to the product. We also calculate the variation of the gas-phase charge distribution (i.e., the partial charges on the QM atoms) along the reaction path in the gas phase by EE-MCMM, and we evaluate the response of the gas-phase partial charges and energy to the electrostatic potential distribution through second order in the Taylor series. Then, we apply the EE-MCMM method to the same reaction in solution, where we use the geometries and electrostatic potentials calculated by the RISM-SCF method[55–57] to compare full RISM-SCF calculations to results predicted by EE-MCMM calculations with all the electrostatic potentials at the Shepard points equal to zero. We employ the same Shepard points as in the gas phase. After the reliability of the EE-MCMM is checked, in the case that only the electrostatic potential is changed, we compare the PES of EE-MCMM calculations to full high-level calculations along an aqueous-solution reaction path. Note that, when we talk about the PES in a liquid-phase solution, we are referring to the electrostatically embedded electronic energy (including nuclear repulsion) of the QM subsystem. The variation of the charge distribution along the reaction path in the aqueous solution is also computed.

The organization of the article is as follows. In the next section, we describe the theoretical methods employed here. The computational details of the EE-MCMM calculations are given in section 3. In section 4, we present the results of the calculations, and the conclusions are summarized in section 5.

2. Theoretical Method

In QM/MM methods, the potential energy is represented as the sum of three terms:

$$V^{\text{total}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) = V^{\text{QM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) + V^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) + V^{\text{MM}}(\mathbf{R}^{\text{MM}}) \quad (2)$$

where $\mathbf{R}$ and $\mathbf{R}^{\text{MM}}$ stand for the collection of the coordinates $\mathbf{R}_a$ and $\mathbf{R}_A{}^{\text{MM}}$ of atoms in the QM and MM regions, respectively, where $a = 1, 2, \ldots, n_1$ and $A = 1, 2, \ldots, n_2$. Here, the first term is the electronic energy of the QM region, $V^{\text{QM}} = \langle\Psi|\hat{H}_0|\Psi\rangle$, with Ψ being the electronic wave function and $\hat{H}_0$ the electronic Hamiltonian (including nuclear repulsions) of the QM region. Note that, although $\hat{H}_0$ depends only on $\mathbf{R}$, Ψ depends on $\mathbf{R}^{\text{MM}}$ as well as $\mathbf{R}$ through $V^{\text{QM/MM}}$. The last term in eq 2 is the MM potential energy function. The QM/MM interaction term $V^{\text{QM/MM}}(\mathbf{R},\mathbf{R}^{\text{MM}})$ can be separated into three terms:

$$V^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) = V_{\text{ele}}^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) + V_{\text{vdW}}^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) + V_{\text{val}}^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) \quad (3)$$

where $V_{\text{ele}}{}^{\text{QM/MM}}$, $V_{\text{vdW}}{}^{\text{QM/MM}}$, and $V_{\text{val}}{}^{\text{QM/MM}}$ are the electrostatic, van der Waals, and valence interaction energies, respectively. Of these three terms, only $V_{\text{ele}}{}^{\text{QM/MM}}$ depends on Ψ. We define the sum of the Ψ-dependent terms, V^{QM} and $V_{\text{ele}}{}^{\text{QM/MM}}$, as the electrostatically embedded QM energy:

$$V^{\text{EEQM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) \equiv V^{\text{QM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) + V_{\text{ele}}^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) \quad (4)$$

The objective of the present study is to reproduce this $V^{\text{EEQM}}(\mathbf{R},\mathbf{R}^{\text{MM}})$ by the EE-MCMM method. Note that V^{EEQM} is called the PES.

We adopt a site–site representation of the QM/MM electrostatic interaction:[55,108–112]

$$V_{\text{ele}}^{\text{QM/MM}}(\mathbf{R}, \mathbf{R}^{\text{MM}}) = \langle\Psi|\hat{\mathbf{Q}}^{\text{T}}\mathbf{\Phi}|\Psi\rangle \quad (5)$$

where $\hat{Q}_a$ is the population operator that generates the partial charge Q_a on the QM atomic site a:

$$Q_a = \langle\Psi|\hat{Q}_a|\Psi\rangle \quad (6)$$

and Φ_a is the electrostatic potential from the MM region:

$$\Phi_a = \sum_{A=1}^{n_2} \frac{Q_A^{\text{MM}}}{|\mathbf{R}_a - \mathbf{R}_A^{\text{MM}}|} \quad (7)$$

where $Q_A{}^{\text{MM}}$ is the effective charge of MM atom A. Note that $\mathbf{Q}$ and $\mathbf{\Phi}$ are n_1-dimensional vectors, and $\mathbf{R}_\text{a}$ and $\mathbf{R}_A{}^{\text{MM}}$ are three-dimentional vectors. By adopting this representation, we can regard V^{EEQM} as a function of $\mathbf{R}$ and $\mathbf{\Phi}$:

$$V^{\text{EEQM}}(\mathbf{R}, \mathbf{\Phi}) = \langle\Psi|\hat{H}_0 + \hat{\mathbf{Q}}^{\text{T}}\mathbf{\Phi}|\Psi\rangle \quad (8)$$

where $\mathbf{R}$ is a $3n_1$-dimensional vector. At this stage, we can extend the MCMM method[75] to the EE-MCMM one straightforwardly.

As in the MCMM method, the potential energy in EE-MCMM is the lowest eigenvalue of a 2 × 2 diabatic Hamiltonian matrix:

$$\mathbf{V}^{\text{EE-MCMM}}(\mathbf{q}, \mathbf{\Phi}) = \begin{pmatrix} V_{11}(\mathbf{q}, \mathbf{\Phi}) & V_{12}(\mathbf{q}, \mathbf{\Phi}) \\ V_{12}(\mathbf{q}, \mathbf{\Phi}) & V_{22}(\mathbf{q}, \mathbf{\Phi}) \end{pmatrix} \quad (9)$$

where we use nonredundant or redundant internal coordinates[113] $\mathbf{q}$ to represent the nuclear coordinates of the QM subsystem. We evaluate $V^{\text{EE-MCMM}}$ and its derivatives in terms of the internal coordinates $\mathbf{q}$; then, we transform the derivatives to the Cartesian coordinate system $\mathbf{R}$. The strategy to be developed involves evaluating a second-order Taylor expression of $V^{\text{EE-MCMM}}$ around a set of interpolation nodes $(\mathbf{R}^{(k)}, \mathbf{\Phi}^{(k)})$, where $k = 1, 2, \ldots, N$, then converting[114] these expansions, for given V_{11} and V_{22}, to second-order expansions of $V_{12}{}^2$ around the interpolation nodes (called Shepard points), and finally evaluating $V_{12}{}^2$ at any arbitrary geometry by Shepard interpolation[85,86] of these expressions.

The lowest eigenvalue of eq 9 is given by

$$V^{\text{EE-MCMM}}(\mathbf{q}, \mathbf{\Phi}) = \frac{1}{2}\Big([V_{11}(\mathbf{q}, \mathbf{\Phi}) + V_{22}(\mathbf{q}, \mathbf{\Phi})] - \big\{[V_{11}(\mathbf{q}, \mathbf{\Phi}) - V_{22}(\mathbf{q}, \mathbf{\Phi})]^2 - 4V_{12}(\mathbf{q}, \mathbf{\Phi})^2\big\}^{\frac{1}{2}}\Big) \quad (10)$$

where $V_{11}(\mathbf{q},\mathbf{\Phi})$ and $V_{22}(\mathbf{q},\mathbf{\Phi})$ are analytic functions that describe V^{EEQM} in the regions of reactants and products. $V_{12}(\mathbf{q},\mathbf{\Phi})$ is evaluated by Shepard interpolation[85,86] as follows:[75]

$$V_{12}(\mathbf{q},\mathbf{\Phi})=\sum_{k=1}^{N}W_k(\mathbf{q},\mathbf{\Phi})V_{12}'(\mathbf{q},\mathbf{\Phi};k) \tag{11}$$

where $W_k(\mathbf{q},\mathbf{\Phi})$ is a normalized weight function:

$$[V_{12}'(\mathbf{q},\mathbf{\Phi};k)]^2=[V_{12}(\mathbf{q},\mathbf{\Phi};k)]^2\,u(\mathbf{q},\mathbf{\Phi};k) \tag{12}$$

where

$$u(\mathbf{q},\mathbf{\Phi};k)=\begin{cases}\exp\left(-\delta/[V_{12}(\mathbf{q},\mathbf{\Phi};k)]^2\right) & [V_{12}(\mathbf{q},\mathbf{\Phi};k)]^2>0\\ 0 & [V_{12}(\mathbf{q},\mathbf{\Phi};k)]^2\le 0\end{cases} \tag{13}$$

and δ is a parameter (we used a very small value of δ, in particular, $(1\times10^{-8})E_{\mathrm{h}}^2$, where $E_{\mathrm{h}}\equiv 1$ hartree), and

$$[V_{12}(\mathbf{q},\mathbf{\Phi};k)]^2=D^{(k)}\Bigg[1+\left(\mathbf{b}_{\mathbf{q}}^{(k)^{\mathrm{T}}}\mathbf{b}_{\mathbf{\Phi}}^{(k)^{\mathrm{T}}}\right)\begin{pmatrix}\Delta\mathbf{q}^{(k)}\\ \Delta\mathbf{\Phi}^{(k)}\end{pmatrix}+\frac{1}{2}\left(\Delta\mathbf{q}^{(k)^{\mathrm{T}}}\Delta\mathbf{\Phi}^{(k)^{\mathrm{T}}}\right)\begin{pmatrix}\mathbf{c}_{\mathbf{qq}}^{(k)} & \mathbf{c}_{\mathbf{q\Phi}}^{(k)}\\ \mathbf{c}_{\mathbf{\Phi q}}^{(k)} & \mathbf{c}_{\mathbf{\Phi\Phi}}^{(k)}\end{pmatrix}\begin{pmatrix}\Delta\mathbf{q}^{(k)}\\ \Delta\mathbf{\Phi}^{(k)}\end{pmatrix}\Bigg] \tag{14}$$

and

$$\Delta\mathbf{q}^{(k)}=\mathbf{q}-\mathbf{q}^{(k)} \tag{15}$$

and

$$\Delta\mathbf{\Phi}^{(k)}=\mathbf{\Phi}-\mathbf{\Phi}^{(k)} \tag{16}$$

For $k = 1, 2, \ldots, N$, the Taylor series coefficients, $D^{(k)}$, $\mathbf{b_q}^{(k)}$, $\mathbf{b_{qq}}^{(k)}$, $\mathbf{c_q}^{(k)}$, $\mathbf{c_{q\Phi}}^{(k)}$, $\mathbf{c_{\Phi q}}^{(k)}$, and $\mathbf{c_{\Phi\Phi}}^{(k)}$, are determined to reproduce V^{EEQM} in eq 8 and its first and second derivatives with respect to $\mathbf{q}$ and $\mathbf{\Phi}$ at the Shepard point $(\mathbf{q}^{(k)},\mathbf{\Phi}^{(k)})$. The expressions for the elements $D^{(k)}$, $\mathbf{b_q}^{(k)}$, and $\mathbf{C_{qq}}^{(k)}$ are given in refs 75 and 82. The other elements are obtained similarly. It is notable that EE-MCMM is the same as the original MCMM in the case when $\mathbf{\Phi} = 0$ and all $\mathbf{\Phi}^{(k)}$ (for $k = 1, 2, \ldots, N$) are also 0.

To implement the above procedure, we need the derivatives of electronic structure calculations of $V^{\mathrm{EEQM}}(\mathbf{R},\mathbf{\Phi})$ with respect to $\mathbf{\Phi}$ in addition to those with respect to $\mathbf{R}$. The first derivative of $V^{\mathrm{EEQM}}(\mathbf{R},\mathbf{\Phi})$ with respect to a component of $\mathbf{\Phi}$ is given by[110]

$$\frac{\partial V^{\mathrm{EEQM}}}{\partial\Phi_a}=\langle\Psi|\hat{Q}_a|\Psi\rangle=Q_a \tag{17}$$

Then, the second partial derivatives of $V(\mathbf{q},\mathbf{\Phi})$ are

$$\frac{\partial^2 V^{\mathrm{EEQM}}}{\partial\Phi_a\,\partial\Phi_b}=\frac{\partial Q_a}{\partial\Phi_b}\equiv\chi_{ab} \tag{18}$$

and

$$\frac{\partial^2 V^{\mathrm{EEQM}}}{\partial\Phi_a\,\partial R_b}=\frac{\partial Q_a}{\partial R_b}\equiv\kappa_{ab} \tag{19}$$

These variables, x_{ab} and κ_{ab}, are known as charge response kernels (CRKs); they describe the QM charge fluctuations due to the external electrostatic potential (which, in applications, will represent the electrostatic effect of the MM region) and to the displacements of the QM atoms. The CRKs x_{ab} and K_{ab} were introduced by Morita and Kato[110,111] and Lu and Yang,[37] respectively. Since these effects are usually not included in MM potential energy functions, we define

$$V_{ii}(\mathbf{q},\mathbf{\Phi})=V_{ii}^{\mathrm{MM}}(\mathbf{q})+V_{ii}^{\mathrm{CRK}}(\mathbf{q},\mathbf{\Phi}) \tag{20}$$

where $V_{ii}{}^{\mathrm{MM}}$ is the MM potential energy function and

$$V_{ii}^{\mathrm{CRK}}(\mathbf{q},\mathbf{\Phi})=\mathbf{Q}^{(i)\mathrm{T}}\Delta\mathbf{\Phi}^{(i)}+\frac{1}{2}\Delta\mathbf{\Phi}^{(i)\mathrm{T}}\boldsymbol{\chi}^{(i)}\Delta\mathbf{\Phi}^{(i)\mathrm{T}}+\Delta\mathbf{\Phi}^{(i)}\kappa^{(i)}\Delta\mathbf{q}^{(i)} \tag{21}$$

where $\mathbf{Q}^{(i)}$, $\kappa^{(i)}$, and $\boldsymbol{\chi}^{(i)}$ are calculated values at the reactant and product, such that the partial charges and CRKs of EE-MCMM agree with electronic structure calculation at the reactant and product, respectively. (Note that the reactant and product correspond to infinitely separated reagents and are not included in the N Shepard points used in eq 11, although we do include the precursor ion−dipole complex and the successor ion−dipole complex.) Then, we can calculate the EE-MCMM potential energy and its derivatives. The calculation steps are the same as those in refs 75 and 82 except that $\mathbf{\Phi}$ is added.

3. Computational Details

We used the MPW1K density functional[115] for the electronic structure calculations on the QM subsystem. The basis set is 6-31G(d,p) for C and H atoms and 6-31+G(d,p) for Cl. We refer to this mixed basis set as 6-31(+)G(d,p). Calculations carried out by direct dynamics, that is, without MCMM or EE-MCMM, will be called direct or full.

Although there can be many choices for the population operator $\hat{Q}_a$, we choose the operator according to Charge Model 4 (CM4).[116] The CM4 charge model is determined from wave-function-dependent charges, the Mayer bond order,[117–119] and empirical parameters that are determined to reproduce experimental or converged theoretical charge-dependent observables:

$$Q_a=Q_a^0+\sum_{b\neq a}B_{ab}(D_{ab}+C_{ab}B_{ab}) \tag{22}$$

where $Q_a{}^0$ is the partial atomic charge from either a Löwdin population analysis (LPA) for nondiffuse basis sets or a redistributed Löwdin population analysis (RLPA) for diffuse basis sets;[120] B_{ab} is the Mayer bond order between atoms a and b, and D_{ab} and C_{ab} are empirical parameters. The RLPA charge is given by

$$Q_a^0(\mathrm{RLPA})=Q_a^0(\mathrm{LPA})+Z_aY_a\sum_{b\neq a}\exp\left(-\alpha_aR_{ab}^2\right)-\sum_{b\neq a}Z_bY_b\exp\left(-\alpha_bR_{ab}^2\right) \tag{23}$$

where Z_a is an empirical parameter, Y_a is the Löwdin population that is associated with the diffuse basis functions on atom a, and α_a is the diffuse orbital exponent on atom a. The Fock matrix and gradient for the Hamiltonian in eq 8 with CM4 charges are given in refs 121 and 122, respectively.

Although the CM4 parameters are available for various density functionals and basis sets, those for the MPW1K/

6-31(+)G(d,p) mixed basis set are unavailable. The reason why we adopted the mixed basis set is that the wave function with MPW1K/6-31G+(d,p) in eq 8 was not converged for $\mathbf{\Phi} \neq 0$ at some geometries. Note that the fixed gas-phase density matrix at a geometry optimized in the gas phase could be used for B_{ab} in the previous study,[122] while this procedure is not appropriate for the present study because the purpose of this study is to describe the global PES. We therefore determined the empirical parameters D_{ab}, C_{ab}, and Z_a for MPW1K/6-31(+)G(d,p) so as to reproduce the CM4 charges obtained with MPW1K/6-31+G(d,p) in the gas phase at three geometries: CH_3Cl, the ion–molecule complex $Cl^-\cdots CH_3Cl$, and the saddle point $[Cl\cdots CH_3\cdots Cl]^-$. The optimized parameters are $D_{ab} = 0.02$ for a C and H pair, $D_{ab} = 0.11$ for a C and Cl pair, and $Z_{ab} = 0.11$ for a Cl atom; the other parameters are set to zero. The mean unsigned error and root-mean-square error of the CM4 charges between MPW1K/6-31+G(d,p) and MPW1K/6-31(+)G(d,p) at the three geometries are 3.6×10^{-3} and 4.9×10^{-3}, respectively. We obtained the Hessian and CRKs by numerical differentiations of the gradients and charges, respectively.

The gas-phase minimum energy path (MEP) was calculated by MCMM by the MC-TINKERATE program.[123] In these calculations, the MEP is the path of steepest descent in mass-scaled coordinates[124] from the saddle point, and the reaction coordinate is the signed distance along the path.

We employed the RISM-SCF method[55–57] to obtain the geometry and electrostatic potential $\mathbf{\Phi}$ on each atom from the MM region in aqueous solution. The reason why we adopted the RISM-SCF method in the present study is that we wanted to check, as a first step, how well the EE-MCMM method can reproduce V^{EEQM} at various geometries and with various electrostatic potential distributions. In the RISM-SCF method, the equilibrium distribution of MM solvent molecules can be calculated in a self-consistent manner. For a fixed subsystem consisting of the solute with coordinates $\mathbf{R}$ and averaging over a subsystem corresponding to the solvent, the FES is approximated as the sum of V^{QM} and the excess chemical potential $\Delta\mu$ coming from solute–solvent interaction:[57]

$$F(\mathbf{R}) = V^{\mathrm{QM}}(\mathbf{R}) + \Delta\mu(\mathbf{R}, \mathbf{Q}) \tag{24}$$

where $\Delta\mu$ is the standard-state free energy of solvation of a solute with fixed geometry $\mathbf{R}$.[59] Note that the FES is another name for a multidimensional potential of mean force.[125,126] This same quantity is also sometimes called[127] the solvent-modified potential energy of the system described by the coordinates $\mathbf{R}$. In the RISM integral equation theory, in conjunction with the hyper-netted chain (HNC) closure relation,[128] $\Delta\mu$ can be expressed as[129]

$$\Delta\mu = -\frac{\rho}{\beta}\sum_a^{n_1}\sum_m^{N_\mathrm{v}}\int_0^\infty \Big[c_{am}(r_{am}) - \frac{1}{2}h_{am}^2(r_{am}) + \frac{1}{2}c_{am}(r_{am})\,h_{am}(r_{am})\Big]4\pi r_{am}^2\,\mathrm{d}r_{am} \tag{25}$$

where r_{am} is the distance between an atom a of the QM solute molecule and an atom m of the MM solvent molecule, $r_{am} = |\mathbf{R}_a - \mathbf{R}_m{}^{\mathrm{MM}}|$, N_v is the number of atoms contained in a solvent molecule ($N_\mathrm{v} = 3$ for water), ρ is the density of the solvent, $\beta = k_\mathrm{B}T$ with k_B being the Boltzmann constant and T the temperature, and c_{am} and h_{am} are the direct and total correlation functions, respectively. Note that c_{am} and h_{am} can be determined from the solute–solvent RISM equation and the HNC closure relation:

$$\tilde{h}_{am}(k_{am}) = \rho^{-1}\sum_b^{n_1}\sum_n^{N_\mathrm{v}} \tilde{w}_{ab}(k_{ab})\,\tilde{c}_{bn}(k_{bn})\,\tilde{H}_{nm}(k_{nm}) \tag{26}$$

and

$$h_{am}(r_{am}) = \exp[-\beta u_{am}(r_{am}) + h_{am}(r_{am}) - c_{am}(r_{am})] - 1 \tag{27}$$

where w_{ab} is the intramolecular correlation function calculated using the QM solute coordinates $\mathbf{R}$ and H_{am} is the pure solvent site density pair correlation function calculated from the solvent−solvent RISM equation; u_{am} is the solute–solvent interaction potential:

$$u_{am}(r_{am}) = \frac{Q_a Q_m^{\mathrm{MM}}}{r_{am}} + 4\epsilon_{am}\left\{\left(\frac{\sigma_{am}}{r_{am}}\right)^{12} - \left(\frac{\sigma_{am}}{r_{am}}\right)^{6}\right\} \tag{28}$$

where ϵ_{am} and c_{am} are the Lennard-Jones parameters and a tilde represents a Fourier transform with wavenumber k_{am} as in

$$\tilde{h}_{am}(k_{am}) = \frac{4\pi}{k_{am}}\int_0^\infty h_{am}(r_{am})\,r_{am}\sin(k_{am}r_{am})\,\mathrm{d}r_{am} \tag{29}$$

With this formalism, V^{QM} and $\mathbf{Q}$ in eq 24 can be determined by eq 8 with

$$\Phi_a = \rho\sum_m^{N_\mathrm{v}}\int_0^\infty \frac{Q_m^{\mathrm{MM}}}{r_{am}}\,g_{am}(r_{am})\,4\pi r_{am}^2\,\mathrm{d}r_{am} \tag{30}$$

where g_{am} is the radial distribution function and

$$g_{am} \equiv h_{am} - 1 \tag{31}$$

We can obtain the self-consistent free energy by iteratively solving eqs 8, 26, and 27 until self-consistency is achieved. The gradient of the free energy F can be calculated analytically.[57]

We optimized the QM geometry on the FES with one or two internal coordinates fixed and then compared V^{EEQM} from the direct calculation (eq 8) to $V^{\mathrm{EE\text{-}MCMM}}$ from the EE-MCMM one (eq 10) at the optimized coordinates and electrostatic potentials. We also calculated the minimum energy path[124] on the FES, and we refer to this as the MFEP. (Since the fixed system in our PMF is a solute, and the averaged subsystem is the solvent, we could also call this an ESP, but we use the more general term throughout the remainder of this article.)

In the RISM-SCF calculation, the Lennard-Jones parameters for the solute atoms were taken from the AMBER force field.[130] The simple point charge model[131] was adopted for solvent–water. The temperature and density of solvent–water were 300 K and 1.0 g/cm,[3] respectively. All of the electronic structure calculations were performed by GAMESS-PLUS[132] based on the GAMESS quantum package,[133] in which we implemented the RISM-SCF routines.

In the MCMM and EE-MCMM calculations, we used a modified MM3 force field[134–136] for the diagonal elements

V_{ii}^{MM} in eq 20. For the bond stretching term, we replaced the MM3 bond stretching function with a Morse[137] potential. The dissociation energy of the Morse function for C−Cl was set equal to 83.7 kcal/mol, which was calculated by MPW1K/6-31(+)G(d,p) and is in good agreement with the experimental value, 83.8 kcal/mol.[138] We also modified the van der Waals energy term as in ref 81; we used the additional parameter $D = 0.01$ in the modified van der Waals energy function. The other parameters are the same as those installed in the TINKER program.[139] We used the same functional form for the normalized weight function as in the original MCMM method[75]

$$W_k(\mathbf{q}) = \frac{\left(\dfrac{1}{d_k(\mathbf{q})}\right)^4}{\displaystyle\sum_{i=1}^{N}\left(\dfrac{1}{d_i(\mathbf{q})}\right)^4} \tag{32}$$

where d_k denotes a generalized distance between $\mathbf{q}$ and $\mathbf{q}^{(k)}$, which is defined as

$$d_k(\mathbf{q}) = \sqrt{\sum_{j=1}^{j_{\max}} (q_j - q_j^{(k)})^2} \tag{33}$$

We employed three bond distances ($j_{\max} + 3$), C−Cl, C−Cl, and Cl−Cl′, to calculate the generalized distance. We did not make the weight function depend on $\mathbf{\Phi}$, although this is possible in principle. All of the EE-MCMM calculations were carried out by the MC-TINKER program,[140] modified for this purpose.

4. Results and Discussion

We applied the new EE-MCMM method to the reaction $\mathrm{Cl^- + CH_3Cl' \rightarrow ClCH_3 + Cl'^-}$ in aqueous solution. The free energy profile of this reaction is much different in aqueous solution from that in the gas phase because the solute–solvent electrostatic interaction at the TS, where there is no dipole moment and the charge is more delocalized, is weaker than that at the reactant. Therefore, this reaction is a good benchmark system for testing the performance of theoretical methods, and consequently various methods have been applied to calculate the free energy profile of this reaction.[29,87–108]

For plotting purposes, we take the difference between two C−Cl distances as the reaction coordinate:

$$z = R_{\mathrm{CCl'}} - R_{\mathrm{CCl}} \tag{34}$$

although the reaction paths along which z and other quantities are computed are the gas-phase MEP and the aqueous-phase MFEP. First, in Figure 1, we compare the gas-phase PES and the aqueous-phase FES with the former evaluated along the direct dynamics MEP and the latter along the direct MFEP. For each curve, the zero of energy corresponds to infinitely separated reagents.

In the gas phase, the ion−dipole complex is 9.7 kcal/mol below reactants, and the potential energy barrier is 3.2 kcal/mol above reactants; both values are in good agreement with experimental values, 10.4[141] and 2.5[142] kcal/mol, respectively. (The best estimate of the gas-phase potential energy barrier is 3.1 kcal/mol.[143]) The ion−dipole complexes are found in the present calculations to be located at $z = \pm 1.378$ Å.

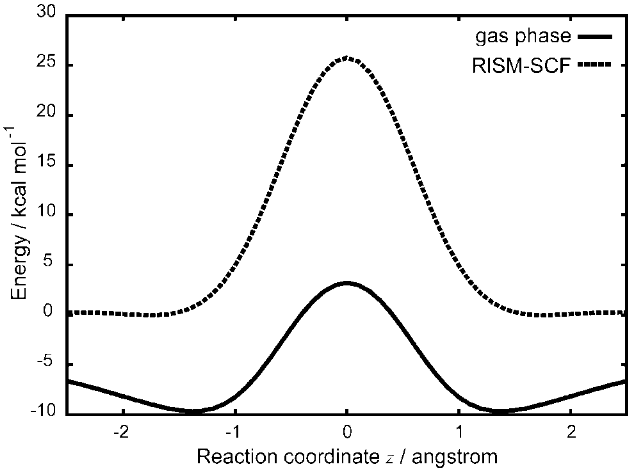


Figure 1. Energy profiles of the $\mathrm{Cl^- + CH_3Cl \rightarrow ClCH_3 + Cl^-}$ reaction: PES profile for gas-phase reaction along the direct MEP (solid) and FES profile for the reaction in aqueous solution along the direct MFEP calculated by RISM-SCF (dashed). Both curves are relative to reactants ($z = -\infty$).

In aqueous solution, the free energy barrier is calculated to be 25.8 kcal/mol, which agrees well with the experimental activation energy, 26.6 kcal/mol.[144] In contrast to the gas-phase reaction, the binding energy for the ion−dipole complex is calculated to be very small. A very shallow minimum (only −0.03 kcal/mol) was found in the FES at $z = 1.744$ Å. Therefore, a practical objective for the EE-MCMM method is to reproduce the potential energy profile for $|z| \leq 1.8$ Å.

4.1. Gas-Phase Reaction. We first constructed a semiglobal potential energy surface in the gas phase using the original MCMM method. The objective region over which we aimed to make this valid was from the reactant ion−dipole complex through the SP to the product ion−dipole complex including the concave side of the reaction path. Note that the previous[75–82] MCMM studies did not attempt to converge the energy surface more than 3/4 of the way down from the barrier, but here we consider the path all the way down to the ion−dipole complexes. The placement of Shepard points was based on the strategy in ref 76, but some modifications were made, as described next.

The first MEP calculation was based on the MCMM-0 surface, which was constructed by electronic structure information at three geometries: the precursor ion−dipole complex, the SP, and the successor ion−dipole complex. (In general, the notation[75,76]MCMM-N' means that the Shepard interpolation is based on Hessians at these three stationary points plus N' nonstationary points.) In the previous studies, we assumed that the V_{11} and V_{22} MM force fields could describe the PES of the local minima in the reactant and product valleys. Therefore, V_{22} was zero for these two points, which will here be called $k = N - 1$ and $k = N$, where $N = N' + 3$. In the present study, we used electronic structure calculations to determine a Taylor series of V_{12}^2 for all N points.

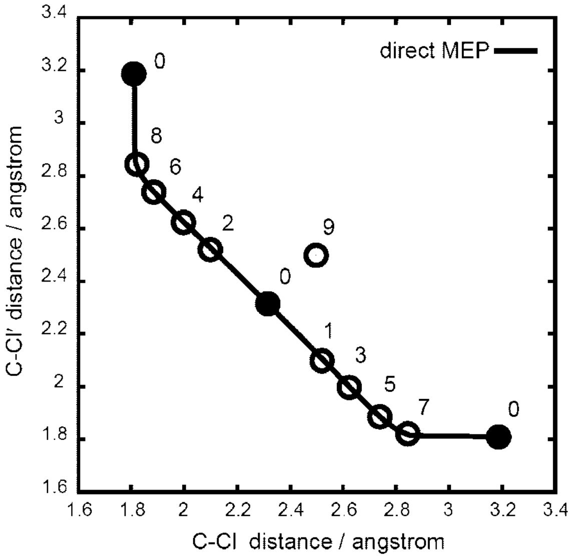


Figure 2. Gas-phase calculations: two-dimensional representation of the direct MEP and the location of Shepard points for the MCMM-9 calculation. Filled circles are stationary points, and open circles are other Shepard points.

In order to keep the symmetry of the reaction, the nonstationary Shepard points were determined at the same time for both the reactant and product sides. We define the energy difference between the ion−dipole complex and the SP as V^*; this is 12.9 kcal/mol for MPW1K/6-31(+)G(d,p) in the gas phase. The first and second supplementary points ($N' = 1, 2$) were taken to be along the MEP of the MCMM-0 run, lower than the SP by 1/4 of V^*. The calculation with these five Shepard points is called MCMM-2 because it involves two supplementary points. The third and fourth supplementary points were taken to be along the MEP of the MCMM-2 run, lower than the SP by 1/2 of V^*. The calculation with these seven Shepard points is called MCMM-4. The fifth and sixth supplementary points were taken to be along the MEP of the MCMM-4 run, lower than the SP by 3/4 of V^*. This calculation is called MCMM-6. The seventh and eighth supplementary points were taken to be along the MEP of the MCMM-6 run, lower than the SP by 7/8 of V^*. This calculation is called MCMM-8. We could connect from the SP to the reactant and product ion−dipole complex smoothly by the MCMM-8 MEP. To reproduce the PES on the concave side of the reaction path, a ninth supplementary point was taken to be located halfway in Cartesian coordinates along a line that connects the reactant ion−dipole complex with the product ion−dipole complex. The calculation including this point is called MCMM-9.

Therefore, we used the electronic structure information at 12 Shepard points (if we consider the symmetry, the number is reduced to seven). The locations of the Shepard points and the direct MEP are shown in Figure 2. It is noted that the purpose of this study is not to reduce the number of Shepard points but to reproduce the semiglobal PES in aqueous solution by EE-MCMM. It is possible to reduce the number of Shepard points by adjusting the force field parameters[81] or changing the strategy for where the Shepard points are placed.

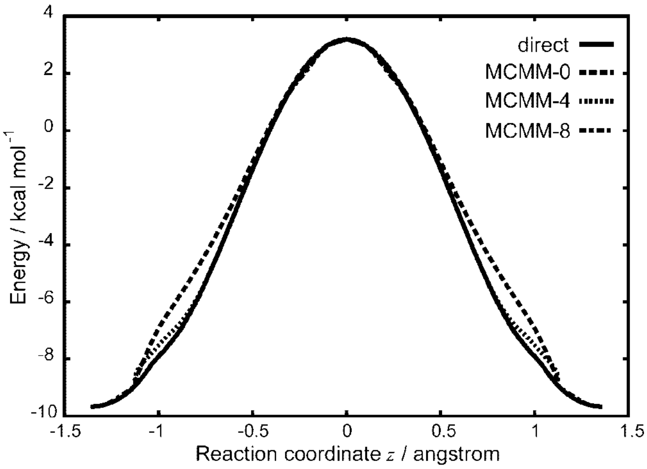


Figure 3. Gas-phase potential energy profiles along the MEP as a function of the reaction coordinate z: direct (solid line), MCMM-0 (dashed line), MCMM-4 (dotted line), and MCMM-8 (dot-dashed line). The dot-dashed line is almost completely hidden by the solid one. All curves are plotted for the direct MEP.

The potential energy profiles of the direct, MCMM-0, MCMM-4, and MCMM-8 gas-phase calculations are shown in Figure 3. The ends of the curves correspond to the precursor and successor ion−dipole complexes. The potential energies of the MCMM-0 and MCMM-4 calculations noticeably differ from the direct one, while the MCMM-8 potential curve is in good agreement with the direct one from the SP all the way to the ion−dipole complexes.

We present equipotential contour plots of the gas-phase PES determined in the MCMM-9 calculation in Figure 4a. The length of the forming C−Cl bond and the breaking C−Cl′ bond are taken as the axes. The remaining coordinates are optimized by direct calculations. Equipotential contour plots of the difference between the MCMM-9 and direct PESs, $V^{\mathrm{MCMM}} - V^{\mathrm{QM}}$, are shown in Figure 4b. In a wide swath from the precursor complex through the SP to the successor complex, including the concave side of the reaction path, the MCMM-9 PES agrees with the direct one within 1 kcal/mol. Therefore, this MCMM-9 PES is accurate enough for dynamics calculations.

The matrix elements of the electronically diabatic Hamiltonian $\mathbf{V}^{\mathrm{MCMM}}$ and the lowest eigenvalue V^{MCMM} are plotted in Figure 5 along four distinguished paths: the path with $R_{\mathrm{CCl}} + R_{\mathrm{CCl'}} = 4.6$ Å (Figure 5a) which goes through the SP, the path with $R_{\mathrm{CCl}} + R_{\mathrm{CCl'}} = 5.0$ Å (Figure 5b) which goes through the reactant and product ion−dipole complexes, the path with $R_{\mathrm{CCl}} = 1.8$ Å which goes through the reactant ion−dipole complex (Figure 5c), and the path with $R_{\mathrm{CCl}} = 2.3$ Å which goes through the SP (Figure 5d). The remaining coordinates are optimized by direct calculations. The matrix element V_{12} has a maximum at the SP and then decreases toward the reactant and product ion−dipole complexes.

To investigate the variation of the partial atomic charges along the reaction path, we carried out an EE-MCMM-9 calculation using the electronic structure information at the same Shepard points as those of MCMM-9. This means that all $\mathbf{\Phi}_a^{(k)}$ values are zero for this EE-MCMM calculation. The partial charges can be obtained by calculating the

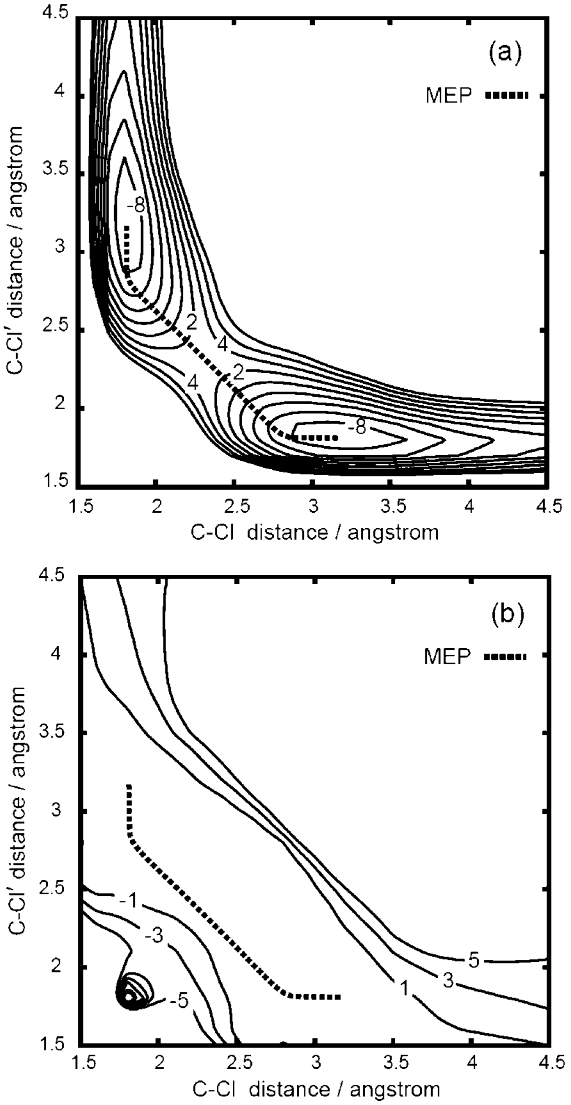


Figure 4. (a) Equipotential contours of the gas-phase PES calculated by MCMM-9. Contour labels are in kcal/mol. Countours are spaced from –8 to +8 by 2 kcal/mol. The zero of energy is at infinitely separated reagents. (b) Equipotential contours of the difference between the gas-phase PESs calculated by the MCMM-9 and direct methods. Contours are spaced from –5 to +5 by 2 kcal/mol.

derivative of $V^{\text{EE-MCMM}}$ in eq 10 with respect to $\mathbf{\Phi}$ as in eq 17, which yields

$$Q_a = \frac{\partial V^{\text{EE-MCMM}}}{\partial \mathbf{\Phi}_a} = \frac{1}{2}\left\{ \frac{\partial V_{11}}{\partial \mathbf{\Phi}_a} + \frac{\partial V_{22}}{\partial \mathbf{\Phi}_a} - \left[\frac{4V_{12}\dfrac{\partial V_{12}}{\partial \mathbf{\Phi}_a} + [V_{11} - V_{22}]\left[\dfrac{\partial V_{11}}{\partial \mathbf{\Phi}_a} - \dfrac{\partial V_{22}}{\partial \mathbf{\Phi}_a}\right]}{[(V_{11} - V_{22})^2 + 4V_{12}]^{1/2}} \right] \right\} \tag{35}$$

Note that the gas-phase charges correspond to evaluating this derivative with all $\mathbf{\Phi}_a = 0$. The partial charges on each atom in the EE-MCMM-9 and direct calculations along each MEP are presented in Figure 6. By construction, the partial charges obtained by eqs 17 and 35 agree exactly at Shepard points, but the figure shows that the changes of the partial charges in the MCMM-9 calculation are quite similar to those in the direct calculation along the whole reaction path. In both cases, the charges of two Cl atoms change significantly along the MEP.

4.2. Reaction in Aqueous Solution. Now we consider the PES for the reaction in aqueous solution; in particular, we will compare $V^{\text{EE-MCMM}}$ to the electrostatically embedded QM energy V^{EEQM}.

When we apply the EE-MCMM method to a reaction in the condensed phase, where $\mathbf{\Phi} \neq 0$, we have to consider how the locations of the Shepard points ($\mathbf{q}^{(k)}$,$\mathbf{\Phi}^{(k)}$) are determined. In general, it is desirable to select the Shepard points so as to make $\Delta\mathbf{q}^{(k)}$ and $\Delta\mathbf{\Phi}^{(k)}$ as small as possible during the statistical sampling in the simulation of the target QM/MM system because EE-MCMM is based on second-order expansions. Several strategies can be considered. One of the strategies, in analogy to the QM/MM-MFEP procedure of Yang and co-workers,[61] is to take the Shepard points along the QM/MM MFEP determined from the potential of mean force in the QM degrees of freedom. In this scheme, the QM geometry and charge distribution are fixed during a MD simulation, then the QM geometry is optimized using the average electrostatic potential and force from the MM atoms; this procedure is repeated until self-consistency between the QM and MM regions is achieved. If the ensemble of reaction paths was restricted to paths that lie close to the MFEP, then this kind of MFEP procedure would always make $\Delta\mathbf{q}^{(k)}$ and $\Delta\mathbf{\Phi}^{(k)}$ small. A drawback to this scheme is that the computational cost of the MFEP calculation is not low. If we were to take supplementary Shepard points along the MFEP of a previous EE-MCMM calculation with fewer Shepard points (as was done in the original MCMM method), hundreds of MD simulation runs would be required, which is undesirable. Furthermore, one expects significant contributions to the reaction rates from paths that differ appreciably from the MFEP.[38–48]

Therefore, we adopted a different strategy for the location of the Shepard points in condensed-phase reactions. We first select Shepard points for a gas-phase reaction in the same way as in the original MCMM method, and then these Shepard points are applied to the reaction in aqueous solution. In other words, all of the Shepard points have $\mathbf{\Phi}^{(k)} = 0$. This means that, as far as the terms relating to the electrostatic potential distribution are concerned, the Taylor series is reduced to a Maclaurin series, or—stated another way—we are using only gas-phase information as input to the Shepard interpolation for the aqueous-phase calculations. We adopted this simple strategy because it has been shown[111] that the linear response relation between $\mathbf{Q}$ and $\mathbf{\Phi}$ (see below), that is, a second-order expansion of V^{EEQM} with respect to $\mathbf{\Phi}$, generally holds well even if the components of $\Delta\mathbf{\Phi}$ become quite large. On the basis of this result, we first generated a semiglobal PES in the gas phase, and then we applied it to the reaction in aqueous solution. It is noted that the computational cost of this strategy is much lower than using a MFEP calculation since only QM gas-phase calculations on the solute are required during the stage of finding the reaction path. Although the present reaction was treated using only eight supplementary points near the gas-

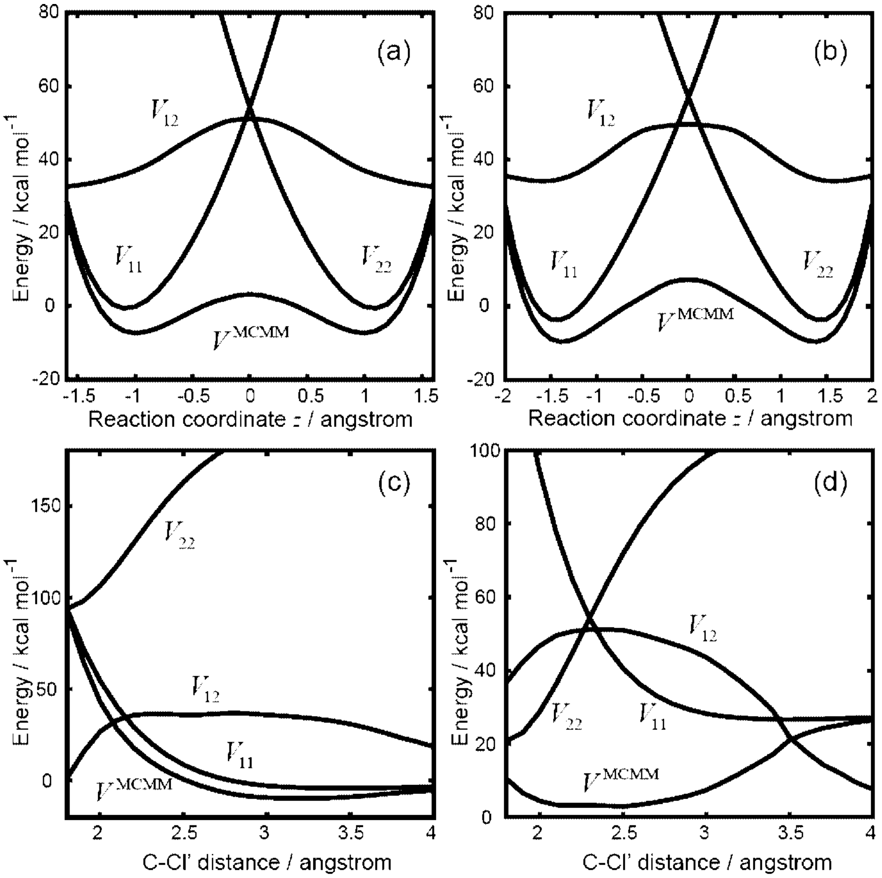


Figure 5. The matrix elements of the electronically diabatic Hamiltonian V^{MCMM} and the lowest eigenvalue V^{MCMM} along the paths with (a) $R_{\text{CCl}} + R_{\text{CCl}'} = 4.6$ Å, (b) $R_{\text{CCl}} + R_{\text{CCl}'} = 5.0$ Å, (c) $R_{\text{CCl}} = 1.8$ Å, and (d) $R_{\text{CCl}} = 2.3$ Å.

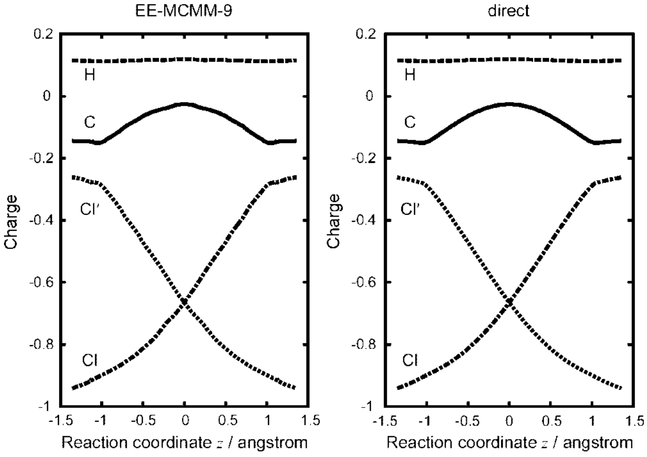


Figure 6. Partial charge on each atom in the EE-MCMM-9 (left) and direct (right) calculations: partial charge on C (solid line), H (dashed line), Cl′ (dotted line), and Cl (dot-dashed line).

phase reaction path and one point off the path, other reactions may require more points off the reaction path. On the other hand, one might be able to use fewer points near the reaction path if their locations are optimized. Further experience will be helpful in understanding these issues.

We first considered the case of $\Delta\mathbf{q} = 0$ and $\Delta\mathbf{\Phi} \neq 0$ to check the reliability. We used the RISM-SCF method to calculate the electrostatic potential on each atom of the solute in aqueous solution at the gas-phase precursor ion−dipole complex and the gas-phase SP. The calculated electrostatic potential distribution is given in Table 1. The electrostatic potential on the Cl ion is larger than those on other atoms at the gas-phase ion−dipole complex because Cl^- has a considerable localized negative charge. In contrast, the electrostatic potential distribution is more uniform at the SP because the charge is more delocalized. We then compared the electrostatically embedded energies and charges of EE-MCMM (calculated with the $\mathbf{\Phi}$ of Table 1) to those of a direct calculation. We also compared the results with those calculated by the original CRK method,[110,111]

Table 1. Electrostatic Potential (in Volts) on Each Atom in Aqueous Solution by RISM-SCF at the Gas-Phase Ion−Dipole Complex and the Gas-Phase Saddle Point

	ion−dipole complex	saddle point
Φ_{C}	4.467	4.753
Φ_{H}	4.596	4.611
$\Phi_{\text{Cl}'}$	3.552	5.211
Φ_{Cl}	7.048	5.211

$$\mathbf{Q}(\text{CRK}) = \mathbf{Q}_0 + \boldsymbol{\chi}\mathbf{\Phi} \tag{36}$$

$$V^{\text{EEQM}}(\text{CRK}) = V_0^{\text{QM}} + \mathbf{Q}_0^{\text{T}}\mathbf{\Phi} + \frac{1}{2}\mathbf{\Phi}^{\text{T}}\boldsymbol{\chi}\mathbf{\Phi} \tag{37}$$

where $\mathbf{Q}_0$ are the charges at $\mathbf{\Phi} = 0$, and V_0^{QM} is the value of $\langle\Psi_0|\hat{H}_0|\Psi_0\rangle$, where Ψ_0 is the gas-phase wave function. The difference between $\langle\Psi|\hat{H}_0|\Psi\rangle$ and $\langle\Psi_0|\hat{H}_0|\Psi_0\rangle$ is accounted for by using the coefficient of 1/2 in the last term of eq 37. Note that the original CRK method and our method

Table 2. Partial Charges (in Units of e) and Electrostatically Embedded QM Energy (in kcal/mol) in the Gas Phase and in Aqueous Solution

	gas phase	solution phase		
	direct	direct	EE-MCMM	Original CRK
		Ion−Dipole Complex[a]		
Q_C	−0.1447	−0.1570	−0.1558	−0.1480
Q_H	0.1157	0.1117	0.1109	0.1110
Q_{Cr}	−0.2559	−0.2036	−0.2072	−0.2093
Q_{Cl}	−0.9425	−0.9744	−0.9699	−0.9751
V^{EEQM}	−9.67	−163.82	−163.72	−163.76
		Saddle Point[a]		
Q_C	−0.0260	−0.0157	−0.0153	−0.0155
Q_H	0.1185	0.1265	0.1264	0.1263
Q_{Cl} (= $Q_{Cl'}$)	−0.6448	−0.6820	−0.6819	−0.6817
V^{EEQM}	3.19	−121.86	−121.85	−121.85

[a] Gas-phase geometries.

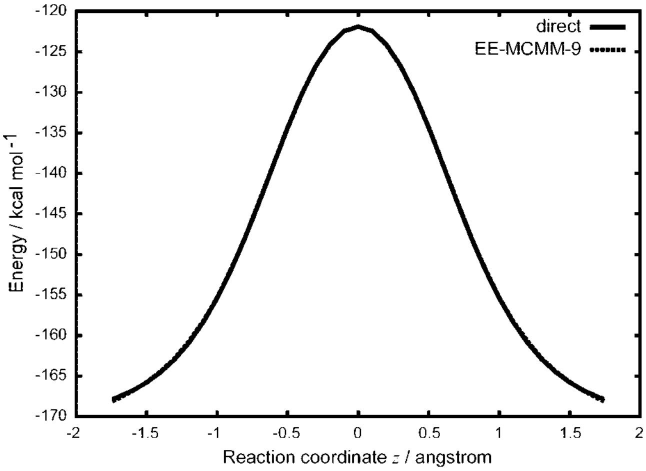


Figure 7. Potential energy profiles along the direct aqueous-phase MFEP: direct RISM-SCF (solid line); EE-MCMM-9 (dashed line).

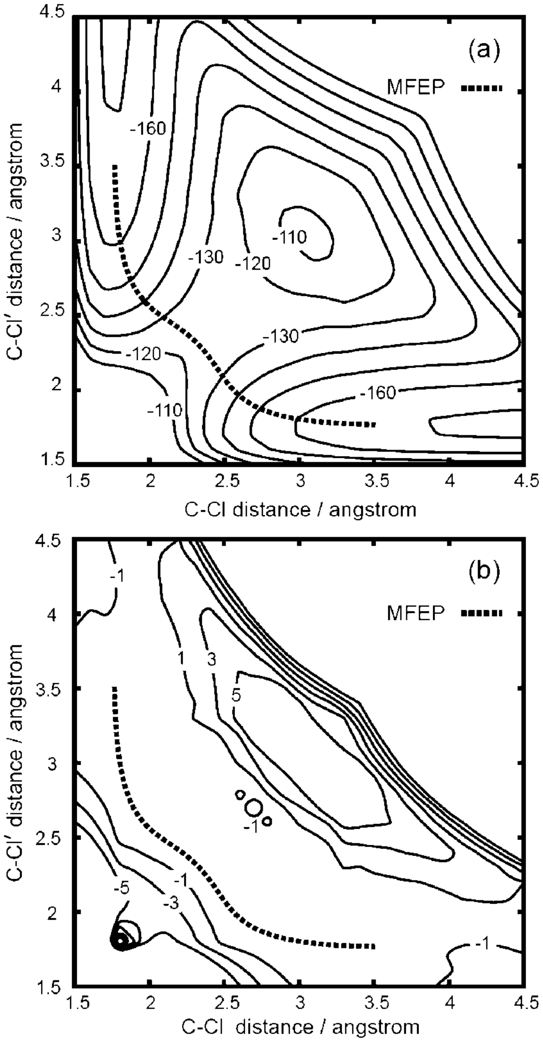


Figure 8. (a) Equipotential contours of the PES calculated by EE-MCMM-9. Contour labels are in kcal/mol. Countours are spaced from –170 to –110 by 10 kcal/mol. (b) Equipotential contours of the difference between the PESs calculated by EE-MCMM-9 and direct methods. Countours are spaced from –5 to +5 by 2 kcal/mol.

differ in the way that the expansion is carried out. The original CRK expands V^{EEQM} itself, while our method expands $V_{12}{}^2$ by using eq 14. The results are shown in Table 2. The aqueous charge distributions obtained by all of the methods in Table 2 are more polarized than the gas-phase charge distribution because of the strong solute–solvent interaction. Both the degrees of charge polarization and the electrostatically embedded energy change upon solution are quite similar in all three methods; the differences are 0.1 kcal/mol or less.

We next calculated the profile of $V^{EE\text{-}MCMM}$ by EE-MCMM-9 along the direct MFEP that was obtained by the RISM-SCF method. The result is presented in Figure 7. The energy is relative to separated reactants in the gas phase. Both edges of the potential energy profiles correspond to the shallow minima of the free energy profile obtained by the RISM-SCF method. The energy difference between the SP and ion−dipole complex is very large compared with gas-phase reaction because of the difference of the solute–solvent interaction. The figure shows that the potential energy profile of EE-MCMM-9 is in very good agreement with that of the direct calculation; in fact, the two curves are essentially on top of one another. We computed equipotential contour plots of $V^{EE\text{-}MCMM}$ as determined in the EE-MCMM-9 calculation; these are shown in Figure 8a. The forming C−Cl bond and the breaking C−Cl′ bond are taken as the axes. The remaining coordinates and the electrostatic potential distribution are optimized by RISM-SCF calculations. Although $V^{EE\text{-}MCMM}$ has a minimum in Figure 8a when both C−Cl distances are increased, neither V^{total} nor F has a minimum in this region. Equipotential contour plots of the difference between the EE-MCMM-9 and direct PESs, $V^{EE\text{-}MCMM} - V^{EEQM}$, are shown in Figure 8b. As in the case of the gas-phase reaction, the EE-MCMM-9 PES agrees with the direct one within 1 kcal/mol in a wide swath from the reactant through the SP to the product, including the concave side of the reaction path. It is notable that we only used electronic structure information of the gas-phase reaction. Nevertheless, we could reproduce the PES for the condensed-phase reaction.

To investigate the effects of the electrostatic potential $\mathbf{\Phi}$ on the matrix elements of the electronically diabatic Hamiltonian $V^{EE\text{-}MCMM}$, we computed these matrix elements along the distinguished path with $R_{CCl} + R_{CCl'} = 4.8$ Å for the

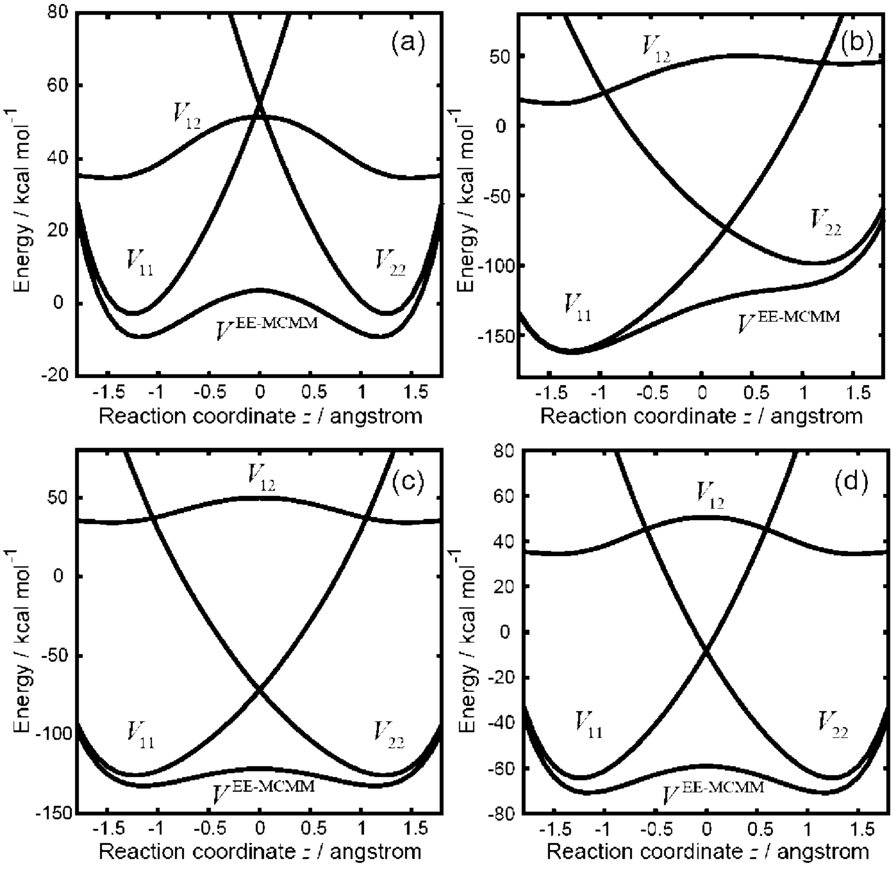


Figure 9. The matrix elements of the electronically diabatic Hamiltonian $V^{\text{EE-MCMM}}$ and the lowest eigenvalue $V^{\text{EE-MCMM}}$ along the path with $R_{\text{CCl}} + R_{\text{CCl}'} = 4.8$ Å for the electrostatic potential distributions with (a) $\mathbf{\Phi} = 0$, (b) $\mathbf{\Phi} = \mathbf{\Phi}^{\text{IDC}}$, (c) $\mathbf{\Phi} = \mathbf{\Phi}^{\text{SP}}$, and (d) $\mathbf{\Phi} = {}^1/_2(\mathbf{\Phi}^{\text{SP}})$.

following four sets of the electronic potential distributions: $\mathbf{\Phi} = 0$ (gas phase), $\mathbf{\Phi} = \mathbf{\Phi}^{\text{IDC}}$, $\mathbf{\Phi} = \mathbf{\Phi}^{\text{SP}}$, and $\mathbf{\Phi} = {}^1/_2\mathbf{\Phi}^{\text{SP}}$, where $\mathbf{\Phi}^{\text{IDC}}$ and $\mathbf{\Phi}^{\text{SP}}$ are the electrostatic potential distribution calculated by RISM-SCF at the gas-phase precursor ion−dipole complex and the gas-phase SP (Table 1). The other remaining coordinates are optimized by direct gas-phase calculations. The results are shown in Figure 9. The diagonal elements V_{11} and V_{22} are strongly stabilized by the external electrostatic potential because the system has negative charge, and all of the values of the electrostatic potential are positive. When $\mathbf{\Phi} = \mathbf{\Phi}^{\text{IDC}}$ (Figure 9b), V_{11} is more stabilized than V_{12} because $\mathbf{\Phi}^{\text{IDC}}$ is favorable to V_{11}. Although the effect of the electrostatic potential on V_{12} is smaller than the effects on V_{11} and V_{12}, the profile of V_{12} with $\mathbf{\Phi} = \mathbf{\Phi}^{\text{IDC}}$ is asymmetric. Therefore, it is important to consider the dependence of V_{12} on external electrostatic potential $\mathbf{\Phi}$.

The charge distribution of the QM subsystem is important in QM/MM calculations since it controls the interaction with the MM subsystem. The partial charge on each atom in the EE-MCMM-9 and direct calculations along the MFEP obtained by the RISM-SCF method is presented in Figure 10. Although there is a slight difference at $|z| > 1.5$ Å, the results of the two calculations are quite similar. Note that no Shepard points were placed at $|z| > 1.378$ Å because the ion−dipole complexes are located at $|z| = 1.378$ Å in the gas phase. If Shepard points are added in such regions, the results will be improved.

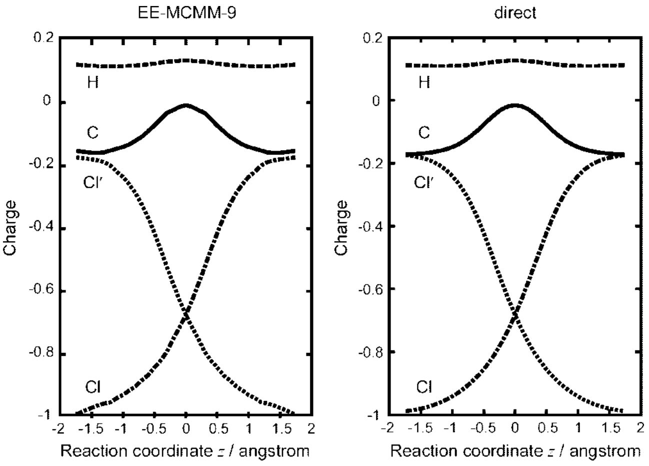


Figure 10. Partial charge on each atom in the EE-MCMM-9 (left) and direct calculations (right) along the MFEP obtained by the RISM-SCF method: partial charge on C (solid line), H (dashed line), Cl′ (dotted line), and Cl (dot-dashed line).

5. Conclusion

In the present work, we proposed a method for generating a potential energy function for a system in the presence of an electrostatic potential. For this purpose, we extended the MCMM method so that the potential energy depends on the electrostatic potential acting on the atomic centers of a subsystem, which is called the QM subsystem. The resulting energy representation can be used to describe PESs defined by a QM/MM method. The charge distribution of the QM subsystem can be obtained by calculating the derivative of the potential energy with respect to the electrostatic potential distribution.

We applied the present method to the degenerate rearrangement $Cl^- + CH_3Cl' \rightarrow ClCH_3 + Cl'^-$ in aqueous solution. We first generated the semiglobal PES in the gas phase by the original MCMM method, and then we generated it in aqueous solution using the same electronic structure information augmented by a Maclaurin series with respect to the electrostatic potential distribution. The calculated potential energy in aqueous solution is very close to that calculated directly without any fitting. The charge distribution in aqueous solution as calculated by the present method is also found to be quite similar to that obtained directly. This shows that we can generate a semiglobal PES in the condensed phase using only electronic structure information in the gas phase. From the perspective of computational cost, it is very efficient that we can use only gas-phase data to determine the location of the Shepard points (in both coordinate space and electrostatic potential distribution space) when we apply the present method to reactions in the condensed phase.

On the basis of the present results, we conclude that the new EE-MCMM method is a very powerful tool for studying reactions in the condensed phase. Although we did not present the results of actual MD simulations here, such applications are now straightforward. An application of the present method to the MD simulation of a condensed-phase reaction is now in progress.

Acknowledgment. This work is supported by the National Science Foundation by grant no. CHE07-04974.

References

(1) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

(2) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.

(3) Gao, J. *Acc. Chem. Res.* **1996**, *29*, 298.

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

(5) Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. *Int. J. Quantum Chem.* **1996**, *60*, 1189.

(6) Truong, T. N.; Truong, T.-T.; Stefanovich, E. V. *J. Chem. Phys.* **1997**, *107*, 1881.

(7) Tongraar, A.; Liedl, K. R.; Rode, B. M. *J. Phys. Chem. A* **1998**, *102*, 10340.

(8) Zhang, Y.; Lee, T.-S.; Yang, W. *J. Chem. Phys.* **1999**, *110*, 46.

(9) Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **1999**, *20*, 1468.

(10) Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1999**, *110*, 10452.

(11) Woo, T. K.; Blöchl, P. E.; Ziegler, T. *J. Phys. Chem. A* **2000**, *104*, 121.

(12) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.

(13) Gogonea, V.; Westerhoff, L. M.; Merz, K. M., Jr *J. Chem. Phys.* **2000**, *113*, 5604.

(14) Chalmet, S.; Rinaldi, D.; Ruiz-Lopez, M. F. *Int. J. Quantum Chem.* **2001**, *84*, 559.

(15) Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. *Theor. Chem. Acc.* **2001**, *105*, 207.

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

(17) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.

(18) Amara, P.; Field, M. J. *Theor. Chem. Acc.* **2003**, *109*, 43.

(19) Vreven, T.; Morokuma, K. *Theor. Chem. Acc.* **2003**, *109*, 125.

(20) Kerdcharoen, T.; Birkenheuer, U.; Krüger, S.; Woiterski, A.; Rösch, N. *Theor. Chem. Acc.* **2003**, *109*, 285.

(21) Nemukhin, A. V.; Grigorenko, B. L.; Topol, I. A.; Burt, S. K. *J. Comput. Chem.* **2003**, *24*, 1410.

(22) Toniolo, A.; Ciminelli, C.; Granucci, G.; Laino, T.; Persico, M. *Theor. Chem. Acc.* **2004**, *111*, 270.

(23) Bathelt, C. M.; Zurek, J.; Mulholland, A. J.; Harvey, J. N. *J. Am. Chem. Soc.* **2005**, *127*, 12900.

(24) Sundararajan, M.; Hillier, I. H.; Burton, N. A. *J. Phys. Chem. A* **2006**, *110*, 785.

(25) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.

(26) To, J.; Sherwood, P.; Sokol, A. A.; Bush, I. J.; Catlow, C. R. A.; van Dam, H. J. J.; French, S. A.; Guest, M. F. *J. Mater. Chem.* **2006**, *16*, 1919.

(27) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.

(28) Senn, H. M.; Thiel, W. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182.

(29) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1985**, *107*, 154.

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

(31) Kollman, P. A.; Kuhn, B.; Donini, O.; Perakyla, M.; Stanton, R.; Bakowies, D. *Acc. Chem. Res.* **2001**, *34*, 72.

(32) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2003**, *125*, 12035.

(33) Xie, L.; Liu, H.; Yang, W. *J. Chem. Phys.* **2004**, *120*, 8039.

(34) Liu, H.; Lu, Z.; Cisneros, G. A.; Yang, W. *J. Chem. Phys.* **2004**, *121*, 697.

(35) Cisneros, G. A.; Liu, H.; Lu, Z.; Yang, W. *J. Chem. Phys.* **2005**, *122*, 114502.

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

(37) Lu, Z.; Yang, W. *J. Chem. Phys.* **2004**, *121*, 89.

(38) Radkiewicz, J. L.; Brooks, C. L., III *J. Am. Chem. Soc.* **2000**, *122*, 225.

(39) Truhlar, D. G.; Gao, J.; Alhambra, C.; Garcia-Viloca, M.; Corchado, J.; Sanchez, M. L.; Villa, J. *Acc. Chem. Res.* **2002**, *35*, 341.

(40) Truhlar, D. G.; Gao, J.; Garcia-Viloca, M.; Alhambra, C.; Corchado, J.; Sanchez, M. L.; Poulsen, T. D. *Int. J. Quantum Chem.* **2004**, *100*, 1136.

(41) Garcia-Viloca, M.; Poulsen, T. D.; Truhlar, D. G.; Gao, J. *Protein Sci.* **2004**, *13*, 2341.

(42) Zou, P.; Osborn, D. L. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1697.

(43) Garrett, B. C.; Truhlar, D. G. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dystra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; p 67.

(44) Roca, M.; Andrés, J.; Moliner, V.; Tuñón, I.; Bertrán, J. *J. Am. Chem. Soc.* **2005**, *127*, 10648.

(45) Thorpe, I. F.; Brooks, C. L., III *J. Am. Chem. Soc.* **2005**, *127*, 12997.

(46) Claeyssens, F.; Ranaghan, K. E.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. *Chem. Comm.* **2005**, 5068.

(47) Pu, J.; Gao, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3140.

(48) Ruiz-Pernía, J.; Silla, E.; Tuñón, I. *J. Phys. Chem. B* **2006**, *110*, 20686.

(49) Rivail, J.-L.; Rinaldi, D. *Comp. Chem.: Rev. Current Trends* **1996**, *1*, 139.

(50) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(51) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(52) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305.

(53) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(54) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

(55) Ten-no, S.; Hirata, F.; Kato, S. *Chem. Phys. Lett.* **1993**, *214*, 391.

(56) Ten-no, S.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1994**, *100*, 7443.

(57) Sato, H.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1996**, *105*, 1546.

(58) Sato, H. *Understanding Chem. React.* **2003**, *24*, 61.

(59) Chuang, Y.-Y.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **1998**, *70*, 887.

(60) Chuang, Y.-Y.; Radhakrishnan, M. L.; Fast, P. L.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 4893.

(61) Hu, H.; Lu, Z.; Yang, W. *J. Chem. Theory Comput.* **2007**, *3*, 390.

(62) Truhlar, D. G.; Liu, Y.-P.; Schenter, G. K.; Garrett, B. C. *J. Phys. Chem.* **1994**, *98*, 8396.

(63) Galván, I. F.; Martín, M. E.; Aguilar, M. A. *J. Comput. Chem.* **2004**, *25*, 1227.

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

(65) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(66) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 7260.

(67) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Status Solidi B* **2000**, *217*, 41.

(68) Hofacker, G. L. *Z. Naturforsch. A* **1963**, *18*, 607.

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

(70) Garrett, B. C.; Truhlar, D. G. *J. Am. Chem. Soc.* **1979**, *101*, 4534.

(71) Morokuma, K.; Kato, S. In *Potential Energy Surfaces and Dynamics Calculations*; Truhlar, D. G., Ed.; Plenum: New York, 1981, pp 243–264.

(72) Miller, W. H. In *Potential Energy Surfaces and Dynamics Calculations*; Truhlar, D. G., Ed.; Plenum: New York, 1981, pp 265–286.

(73) Truhlar, D. G.; Brown, F. B.; Steckler, R.; Isaacson, A. D. In *The Theory of Chemical Reaction Dynamics*; Clary, D. C., Ed.; Reidel: Dordrecht, The Netherlands, 1986; NATO ASI Series C, 70, pp. 285–329.

(74) Fernandez-Ramos, A.; Ellingson, B. A.; Garrett, B. C.; Truhlar, D. G. *Rev. Comput. Chem.* **2007**, *23*, 125.

(75) Kim, Y.; Corchado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718.

(76) Albu, T. V.; Corchado, J. C.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 8465.

(77) Lin, H.; Pu, J.; Albu, T. V.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4112.

(78) Kim, K. H.; Kim, Y. *J. Chem. Phys.* **2004**, *120*, 623.

(79) Kim, Y.; Kim, Y. *J. Phys. Chem. A* **2006**, *110*, 600.

(80) Lin, H.; Zhao, Y.; Tishchenko, O.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1237.

(81) Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13530.

(82) Tishchenko, O.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 938. The final corrected equations for the gradients and Hessians with respect to coordinates (O. Tishchenko and D. G. Truhlar, to be published) are given in an appendix to the *MC-TINKER* manual, available online at http://comp.chem.umn.edu/mc-tinker/.

(83) Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*, 5048.

(84) Albu, T. V.; Espinosa-García, J.; Truhlar, D. G. *Chem. Rev.* **2007**, *107*, 5101.

(85) Ischtwan, J.; Collins, M. A. *J. Chem. Phys.* **1994**, *100*, 8080.

(86) Nguyen, K. A.; Rossi, I.; Truhlar, D. G. *J. Chem. Phys.* **1995**, *103*, 5522.

(87) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1984**, *106*, 3049.

(88) Bash, P. A.; Field, M. J.; Karplus, M. *J. Am. Chem. Soc.* **1987**, *109*, 8092.

(89) Kozaki, T.; Morihashi, K.; Kikuchi, O. *J. Am. Chem. Soc.* **1989**, *111*, 1547.

(90) Huston, S. E.; Rossky, P. J.; Zichi, D. A. *J. Am. Chem. Soc.* **1989**, *111*, 5680.

(91) Tucker, S. C.; Truhlar, D. G. *J. Am. Chem. Soc.* **1990**, *112*, 3347–3361.

(92) Zhao, X. G.; Tucker, S. C.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 826.

(93) Basilevsky, M. V.; Chudinov, G. E.; Napolov, D. V. *J. Phys. Chem.* **1993**, *97*, 3270.

(94) Mathis, J. R.; Bianco, R.; Hynes, J. T. *J. Mol. Liq.* **1994**, *61*, 81.

(95) Truong, T. N.; Stefanovich, E. V. *J. Phys. Chem.* **1995**, *99*, 14700.

(96) Pomelli, C. S.; Tomasi, J. *J. Phys. Chem. A* **1997**, *101*, 3561.

(97) Cossi, M.; Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1998**, *297*, 1.

(98) Mo, Y.; Gao, J. *J. Comput. Chem.* **2000**, *21*, 1458.

(99) Safi, B.; Choho, K.; Geerlings, P. *J. Phys. Chem. A* **2001**, *105*, 591.

(100) Ohmiya, K.; Kato, S. *Chem. Phys. Lett.* **2001**, *348*, 75.

(101) Gao, J.; Garcia-Viloca, M.; Poulsen, T. D.; Mo, Y. *Adv. Phys. Org. Chem.* **2003**, *38*, 161.

(102) Mo, S. J.; Vreven, T.; Mennucci, B.; Morokuma, K.; Tomasi, J. *Theor. Chem. Acc.* **2004**, *111*, 154.

(103) Vayner, G.; Houk, K. N.; Jorgensen, W. L.; Brauman, J. I. *J. Am. Chem. Soc.* **2004**, *126*, 9054.

(104) Sato, H.; Sakaki, S. *J. Phys. Chem. A* **2004**, *108*, 1629.

(105) Freedman, H.; Truong, T. N. *J. Phys. Chem. B* **2005**, *109*, 4726.

(106) Song, L.; Wu, W.; Hiberty, P. C.; Shaik, S. *Chem.—Eur. J.* **2006**, *12*, 7458.

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

(108) Su, P.; Wu, W.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A*, to be published.

(109) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(110) Morita, A.; Kato, S. *J. Am. Chem. Soc.* **1997**, *119*, 4021.

(111) Morita, A.; Kato, S. *J. Chem. Phys.* **1998**, *108*, 6809.

(112) Hayashi, S.; Ohmine, I. *J. Phys. Chem. B* **2000**, *104*, 10678.

(113) Wilson, E. B., Jr.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; Dover: New York, 1955.

(114) Chang, Y.-T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884.

(115) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811.

(116) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(117) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270.

(118) Mayer, I. *Chem. Phys. Lett.* **1985**, *117*, 396.

(119) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 477.

(120) Thompson, J. D.; Xidos, J. D.; Sonbuchner, T. M.; Cramer, C. J.; Truhlar, D. G. *PhysChemComm* **2002**, 117.

(121) Zhu, T.; Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, *109*, 9117.

(122) Zhu, T.; Li, J.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 5503.

(123) Albu, T. V.; Tishchenko, O.; Corchado, J. C.; Kim, Y.; Villà, J.; Xing, J.; Lin, H.; Truhlar, D. G.;. *MC-TINKERATE*, version 2007; University of Minnesota: Minneapolis, MN, 2007.

(124) Isaacson, A. D.; Truhlar, D. G. *J. Chem. Phys.* **1982**, *76*, 1380.

(125) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

(126) Hill, T. L. *Statistical Mechanics: Principles and Selected Applications*; Dover: New York, 1956; pp 193–194.

(127) Kinoshita, M. *Understanding Chem. React.* **2003**, *14*, 101.

(128) Hirata, F. *Understanding Chem. React.* **2003**, *24*, 1.

(129) Singer, S. J.; Chandler, D. *Mol. Phys.* **1985**, *55*, 621.

(130) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. G.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(131) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. P.; Hermans, J.; Pullman, B. *Intermolecular Forces*; Reidel, Dordrecht, The Netherlands, 1981.

(132) Chamberlin, A. C.; Pu, J.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Zhu, T.; Hawkins, G. D.; Chuang, Y.-Y.; Fast, P. L.; Lynch, B. J.; Liotard, D. A.; Rinaldi, D.; Gao, J.; Cramer, C. J.; Truhlar, D. G.; *GAMESSPLUS*, version 4.8.; University of Minnesota: Minneapolis, MN, 2006.

(133) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(134) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(135) Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8566.

(136) Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8576.

(137) Herzberg, G. *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules*; 2nd ed.; D. Van Nostrand: Princeton, NJ, 1950; p 101.

(138) Linstrom, P. J.; Mallard, W. G.; NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology: Gaithersburg, MD, 2005. http://webbook.nist.gov.

(139) Ponder, J. W.; *TINKER*, version 3.5; Washington University: St. Louis, MO, 1997.

(140) Tishchenko, O.; Albu, T. V.; Corchado, J. C.; Kim, Y.; Villà, J.; Xing, J.; Lin, H.; Truhlar, D. G.; *MC-TINKER*, version 2007; University of Minnesota: Minneapolis, MN, 2007.

(141) Li, C.; Ross, P.; Szulejko, J. E.; McMahon, T. B. *J. Am. Chem. Soc.* **1996**, *118*, 9360.

(142) Wladkowski, B. D.; Brauman, J. I. *J. Phys. Chem.* **1993**, *97*, 13158.

(143) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.

(144) McLennan, D. J. *Aust. J. Chem.* **1978**, *31*, 1897.

CT800004Y

JCTC Journal of Chemical Theory and Computation

Tight-Binding Configuration Interaction (TBCI): A Noniterative Approach to Incorporating Electrostatics into Tight Binding

Mark A. Iron,[†,§] Andreas Heyden,[†,||] Grażyna Staszewska,[†,‡] and Donald G. Truhlar*[,†]

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431, and Institute of Physics, Nicolaus Copernicus University, ul. Grudziądzka 5, 87-100 Toruñ, Poland

Received December 18, 2007

Abstract: We present a new electronic structure approximation called Tight Binding Configuration Interaction. It uses a tight-binding Hamiltonian to obtain orbitals that are used in a configuration interaction calculation that includes explicit charge interactions. This new method is better capable of predicting energies, ionization potentials, and fragmentation charges than the Wolfsberg−Helmholz Tight-Binding and Many-Body Tight-Binding models reported earlier (Staszewska, G.; Staszewski, P.; Schultz, N. E.; Truhlar, D. *Phys. Rev. B* **2005**, *71*, 045423). The method is illustrated for clusters and nanoparticles containing aluminum.

1. Introduction

Tight-binding (TB) theory[1–3] (also called extended Hückel theory) is well-suited to modeling the electronic structure and potential energy surfaces of materials and nanoparticles in that it incorporates quantum mechanical effects with a minimum of computational expense. It is, however, a very approximate theory and can only provide useful accuracy if the matrix elements are empirically parametrized. It is well-known in a variety of contexts that empirically parametrized theories work best when the functional or operational form of the theory incorporates the dominant physical interactions. Tight-binding theory has the functional form of a one-electron eigenvalue problem; this class of functional forms can be derived[4,5] from Kohn−Sham density functional theory[6] by assuming that the densities of the atoms of a given atomic number are all close to a reference density for that atomic number. This assumption breaks down for many problems, in particular those involving bond breaking and formation, variable coordination numbers, multiple oxidation states, and polar or ionic bonds. Thus, the question arises of whether there might be other functional or operational forms that are only slightly more computationally expensive but better incorporate the atomic interactions.

Kohn−Sham theory itself has the form of a pseudoeigenvalue problem in that the Kohn−Sham matrix to be diagonalized depends on the occupied orbitals, which in turn depend on the field produced by the orbital eigenvectors. This kind of problem can be solved by an iterative process, leading to the familiar self-consistent field problem.[7–9] Various versions of tight binding with occupancy-dependent terms in the Hamiltonian (e.g., iterative extended Hückel theory,[10] the Anderson−Newns model,[11] the Hubbard model,[12] the Grimley−Pisani model,[13] self-consistent charge density functional tight binding (SCC-DFTB),[14,15] and the generalized Hubbard model[16–18]) have been proposed, and some of these models have enjoyed considerable success. Nonetheless, occupancy-dependent Hamiltonians require an iterative method for their solution, and an iterative process raises the cost and complexity of the method, especially if one is computing the gradients and/or Hessians[19] for dynamics. Furthermore, there is the added inconvenience that one must develop computational strategies for dealing with unstable cases where the iterative process diverges, stalls, or oscillates. The present article is devoted to trying to solve some of the same issues that motivated iterative tight-binding

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

[†] University of Minnesota.

[‡] Nicolaus Copernicus University.

[§] Current address: Computational Chemistry Unit, Department of Chemical Research Support, Weizmann Institute of Science, Rehovot, Israel 76100.

[||] Current address: Department of Chemical Engineering, University of South Carolina, Columbia, SC 29208.

10.1021/ct700343t CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/25/2008

schemes, especially finding the optimum charge distribution, but with noniterative methods.

In general, tight binding has almost always been considered a way to approximate a single-configuration wave function, such as the ones in Hartree–Fock,[20–24] Hartree–Fock–Slater,[25] Kohn–Sham,[6] or dynamical Hartree–Fock[26] theories. One can make Kohn–Sham theory more accurate by improving the one-electron Hamiltonian (by using better density functionals), and such an approach is analogous in some respects to better parametrization of the tight-binding Hamiltonian, just as the iterative refinement of the Kohn–Sham Hamiltonian is the direct analogue of the iterative tight-binding methods discussed above. However, the more traditional approach in quantum chemistry has been configuration interaction (CI), either by variational theory,[27] by perturbation (or many-body) theory,[28,29] or by coupled cluster theory;[29] in these theories, the wave function is a linear combination of configuration state functions (CSFs), each built from Hartree–Fock orbitals (as in the so-called post-Hartree–Fock correlation methods[30]), from non-self-consistent orbitals (as in valence bond theory[31,32]), or from Kohn–Sham orbitals.[33] This suggested to us that one might improve tight-binding theory by using the tight-binding orbitals to perform an approximate configuration interaction calculation. We propose and test such an approach in the present article.

The main reason why one usually prefers using self-consistent-field orbitals for configuration interaction is that it makes the matrix elements connecting the self-consistent configurations to single excitations all equal to zero.[34–38] Since we are using tight-binding orbitals, which are not self-consistent, we will include single excitations. In fact, the single excitations are the ones most responsible for charge redistribution. Since our main goal is to include charge redistribution effects without resorting to iterative methods, the single excitations are the key element of the new method.

A second motivation for the development of the new method, called tight-binding configuration interaction (TBCI), is that single-configuration methods like tight-binding often lead to incorrect electron distributions upon bond breaking.[39,40] The new method is designed to improve this so that the new tight-binding method is more reasonable in describing bond-breaking dynamical processes.

Aluminum clusters and nanoparticles were chosen as the initial application for evaluating the new method. Recently, we have devoted considerable effort to studying these systems.[39,41–48] Analytical potential energy functions (PEFs) were developed that accurately predict cluster energies.[42,45] A number of TB and many-body TB (MBTB) models were also parametrized.[39,44,46] In addition, when potentials based on these methods were used, the properties of aluminum were explored by molecular dynamics and Monte Carlo calculations.[49,50] There are a number of reasons for the interest in this metal. It is of technological and industrial value as a lightweight, rust-resistant metal, and it is an ingredient for high-energy fuels (e.g., as a component of solid rocket propellant) and potentially as a hydrogen-storage medium.[51] The use of aluminum nanoparticles, instead of bulk metal, is expected to enhance the latter two properties.[51–54] Metal nanoparticles provide a challenging test for approximate electronic structure theories because it has been demonstrated that the properties of nanometer-sized particles can depend dramatically on size, occasionally in an unpredictable manner, as compared to those of bulk metal.[55–65]

2. Theory

First, we present the theory behind the new electronic structure method for a system involving only a single element. Afterward, the extension of the method to heteroatomic systems is described.

The CI wave function is written as

$$\Psi = \sum_j C_j \Phi_j \tag{1}$$

where C_j is a CI coefficient and Φ_j is a CSF. We follow the common practice in TB of using the frozen-core approximation and thus consider only the valence electrons. Each CSF is a Hartree product:

$$\Phi_j = \varphi_1^{o_1^{(j)}} \varphi_2^{o_2^{(j)}} \cdots \varphi_M^{o_M^{(j)}} \tag{2}$$

where φ_k is an orbital, $o_k^{(j)}$ is the occupancy number (0, 1, or 2) of orbital k in CSF j, and M is the number of orbitals. The sum of the occupancy numbers in any CSF equals the number of valence electrons:

$$N_{\mathrm{val}} = \sum_k o_k^{(j)} \tag{3}$$

The one-electron density matrix of the CI wave function is approximated as

$$P_{mn} = \sum_j |C_j|^2 P_{mn}^{(j)} \tag{4}$$

where $P_{mn}{}^{(j)}$ is the density matrix of CSF j. The total CI energy E is the lowest eigenvalue of the CI matrix. Rather than construct and diagonalize the full CI matrix, which is far too large even if we knew or had approximations for the off-diagonal elements, the lowest eigenvalue is approximated using a weighted version of the degeneracy-corrected perturbation theory[66] energy:

$$E = \sum_j |C_j|^2 \Gamma_j \tag{5}$$

$$\Gamma_j = E_j + \frac{1}{2}\sum_i \left[(E_i - E_j) - \sqrt{(E_i - E_j)^2 + 4\chi_{ij}^2}\right] \tag{6}$$

where E_j is the energy of CSF j, and χ_{ij} is the *coupling element* between CSFs i and j, when $i \neq j$, and is a constant called the *diagonal coupling constant* when $i = j$ (the $i = j$ term is included in the sum to ensure size extensivity even in the presence of degeneracies). This form (eqs 5 and 6) was chosen for several reasons: it leads to a size-consistent method, it gives the correct answer for a 2 × 2 CI matrix with known off-diagonal elements, the ground-state energy is below the lowest CSF energy, and the ground-state energy approaches the energy of the lowest CSF in the limit of well-separated CSF energies.

In lieu of eq 5 for determining the TBCI energy, one could consider using

$$E = \sum_j |C_j|^2 E_j \tag{7}$$

However, this equation will lead to total energies that are higher than that of the lowest-energy CSF, whereas eq 5 will always give total energies that, as one finds in a variational calculation, are lower. Another possible formula for the total energy is based on the free energy formula:

$$E = -\Delta \ln\left[\sum_j \exp\left(-\frac{E_j}{\Delta}\right)\right] \tag{8}$$

Equation 8 gives overall energies that are lower than that of the lowest-energy CSF, but it is not size-extensive. We found that the combination of eqs 5 and 24 (*vide infra*) is the most satisfactory formula for the configuration interaction energy, and we did not use eqs 7 or 8.

The remaining tasks are to develop approximations for $|C_j|^2$, χ_{ij}, and E_j.

The molecular orbitals (MOs) are expanded in a minimal basis set $\{\eta_a\}$ of atomic orbitals:

$$\varphi_k = \sum_a c_{ak}\eta_a \tag{9}$$

The atomic orbitals are assumed to be orthonormal. The orbital coefficients c_{ak} are obtained by solving the tight-binding orbital eigenvalue equations, given in matrix form by

$$\mathbf{Hc} = \boldsymbol{\varepsilon}\mathbf{c} \tag{10}$$

where $\mathbf{H}$ is the $M \times M$ tight-binding Hamiltonian matrix (i.e., the one-electron Hamiltonian) with elements H_{ab}, $\mathbf{c}$ is an $M \times M$ matrix in which each column is an eigenvector, and $\boldsymbol{\varepsilon}$ is a diagonal matrix of orbital eigenvalues. The elements of $\mathbf{c}$ are the orbital coefficients in eq 9. On the basis of previous work,[39,44,46] we will assume zero differential overlap (ZDO)[67,68] so that the density matrix of CSF j in the atomic orbital representation becomes

$$P_{ab}^{(j)} = \delta_{ab}\sum_k^{MO} o_k^{(j)}|\, c_{ak}|^2 \tag{11}$$

The ZDO approximation is the reason that the atomic-orbital overlap matrix is neglected in eq 10. In a previous work on TB, it was found that neglecting the overlap gave a more balanced set of approximations.[46]

The energy of a given CSF is given by

$$E_j = E_{\text{val}}^{(j)} + E_{\text{CB}}^{(j)} + V_{\text{rep}} + E^{(0)} \tag{12}$$

The valence energy, $E_{\text{val}}^{(j)}$, is given by

$$E_{\text{val}}^{(j)} = \sum_k o_k^{(j)}\varepsilon_k \tag{13}$$

In some versions[39,46,69] of TB, the valence energy contains an additional term, $\delta_{o_k,2\mu_{\text{penalty}}}$, which introduces a penalty (u_{penalty}) for two electrons occupying the same orbital. This allows for ground states that are not singlets or doublets and for homolytic bond dissociations of singlet molecules. One motivation behind developing the TBCI model is to find a more satisfactory means of allowing for these possibilities.

In eq 12, V_{rep}, is the core−core repulsion, which is assumed to be a function of geometry but not of the CSF (i.e., index j). It is modeled in the present work by a pairwise-additive potential:

$$V_{\text{rep}} = \kappa\sum_A\sum_{B>A}\exp(-\tau\, R_{AB}) \tag{14}$$

where A and B are atom labels, and κ and τ are empirical parameters. This form is similar to that used in our previous TB work[39] where we also included a prefactor $(R_{AB})^{-u_{AB}}$ with u_{AB} being a positive constant parameter; this prefactor was omitted here since the optimized u_{AB} tended to be small, and the prefactor was found in preliminary work to have an insignificant effect. $E^{(0)}$ is a constant, independent of both geometry and CSF, that determines the zero of energy. The charge balance (CB) energy, $E_{\text{CB}}^{(j)}$, is primarily responsible for describing the electrostatics in the system. In order to determine E_j, what remains is to define $\mathbf{H}$, $E_{\text{CB}}^{(j)}$, and $E^{(0)}$.

The Hamiltonian $\mathbf{H}$ is modeled by the Wolfsberg−Helmholz approximation.[39,70] The diagonal elements are

$$H_{aa} = -U_Z^l \tag{15}$$

where U_Z^l is a parameter, different for each type of orbital. Note that these matrix elements depend on the atomic number Z and the orbital angular momentum quantum number l. In our previously published TB models, U_Z^l were taken as valence-state ionization potentials (VSIP or I_Z^l).[39] In TBCI, because of the CI treatment and CB term, one cannot make the same assumption, and the equations used for assigning U_Z^l are given later in this section.

The off-diagonal elements, called either transfer or hopping integrals, are[39,70]

$$H_{ab} = K_{l_a l_b m}\frac{H_{aa} + H_{bb}}{2}S_{ab} \tag{16}$$

where S_{ab} is the atomic-orbital overlap integral and $K_{l_i l_j m}$ is an empirical constant that depends on the orbital angular momenta l_i and angular momentum projection quantum number m on the axis connecting atoms A and B. This approximation was initially proposed by Mulliken[71] and has been used by Hoffmann[2,72–75] and Wolfsberg and Helmholz[70] and in the TB and MBTB models for aluminum.[39] We take each basis function η_a to be a single Slater-type orbital.[76] Then, the overlap matrix elements can be readily calculated. For example, when the orbital exponents on the two centers are equal, one can use the expressions determined by Jones[77]

$$S_{ab} = P_6(\zeta\, R_{AB})\exp(-\zeta\, R_{AB}) \tag{17}$$

where $P_6(x)$ is a sixth-order polynomial, A and B label the atoms on which orbitals a and b are centered, R_{AB} is the interatomic distance, and ζ is the Slater-orbital exponential parameter. This expression is sufficient for Al because fortuitously the s- and p-exponential parameters, as determined in the minimal basis set Hartree−Fock calculations by Clementi and Raimondi,[78] happen to be nearly the same; for the s-p overlap, the average of the two exponential parameters is used, in a manner similar to that of our previous work.[39] Jones has also provided equations for the overlap for the more general case where the two exponential parameters (ζ_a and ζ_b) are different.[79]

As in previous work on TB,[39] various models can be formulated by using different sets of $K_{l_il_jm}$ and $\zeta_{l_il_jm}$. For aluminum, there are four relevant combinations of l_i, l_j, and m: $ss\sigma$, $sp\sigma$ (equivalent to $ps\sigma$), $pp\sigma$, and $pp\pi$. One could optimize either a single $K_{l_il_jm}$ parameter for all four combinations (the Wolfsberg–Helmholz or WH model) or separate $K_{l_il_jm}$ parameters for each permutation (the extended Wolfsberg–Helmholz or EWH model). In the EWH model, we use Clementi and Raimondi's ζ_a exponential parameters (except that for Al, we average the *s*- and *p*-exponential parameters in the $sp\sigma$ case in order to use the simpler expression for the overlap in eq 17, *vide supra*). In the optimized Wolfsberg–Helmholz or OWH model, the four Slater exponential parameters and four separate $K_{l_il_jm}$ parameters are optimized.[39]

In order to calculate the charge balance energy, one must first calculate the partial atomic charges for each CSF. These are determined using a Mulliken–Coulson population analysis,[80–82] which, using eq 11, yields

$$q_A^{(j)} = Z_A^* - \sum_{a\in A} P_{aa}^{(j)} \tag{18}$$

where Z_A^* is the shielded nuclear charge of atom A (i.e., the number of valence electrons in the neutral atom, which is three for aluminum). The notation "$a \in A$" means that the sum in eq 18 is over all orbitals a on center A.

The charge balance energy is written as the sum of on-site and intersite interactions:

$$E_{\mathrm{CB}}^{(j)} = \sum_{A} \sum_{B\geq A} \gamma(R_{AB})\, q_A^{(j)} q_B^{(j)} \tag{19}$$

where $\gamma(R_{AB})$ is a Coulomb integral (*vide infra*) and $R_{AA} \equiv 0$. The diagonal terms in eq 19 are called on-site interactions. These interactions are the key element in Hubbard[12,16–18] and DFT+U[83,84] theories. The present theory also includes the intersite interactions. If the latter are neglected, the method is called TB+U, while the full theory is termed TBCI.

A key difference from SCC-DFTB[14,15] is that the electrostatic terms are added to the energy of each configuration after determining the orbitals, rather than iteratively while determining the orbitals. Their effects are included by the charge balance term of eq 19 and the configuration interaction approximation of eq 5. *Thus, TBCI is a noniterative method.* The slowest step in either SCC-DFTB or TBCI is the diagonalization of the one-electron Hamiltonian matrix in eq 10, which approximately scales as N^3, where N is the number of atoms in the system. Since SCC-DFTB repeats this step in each iterative cycle, whereas TBCI only does this step once, TBCI is significantly faster. Furthermore, the absence of an iterative step makes the analytical gradients particularly stable and precise. Finally, the neglect of the overlap integrals is a further feature that makes TBCI faster.

The on-site interactions $\gamma(R_{AA})$ could be approximated by the Pariser formula:[85]

$$\gamma(R_{AA}) \equiv \gamma_{AA} = \mathrm{IP}_A - \mathrm{EA}_A \tag{20}$$

where IP_A and EA_A are, respectively, the ionization potential and electron affinity of atom A; alternatively, they can be taken as empirical parameters. In the current implementation, the second option was chosen. It is convenient to use atomic units and write

$$\gamma_{AA} = \frac{e^2}{\alpha_A} \tag{21}$$

where e is the charge of an electron (unity in atomic units), and α_A is a parameter with units of length that will be called the *atomic Coulomb radius*.

The intersite charge interaction terms can be approximated in various ways. For example, one possible approximation for the Coulomb integral is based on a dielectric screening model[86,87] yielding, in atomic units

$$\gamma(R_{AB}) = e^2\left[\alpha_A\alpha_B \exp\left(\frac{-R_{AB}^2}{d\alpha_A\alpha_B}\right) + R_{AB}^2\right]^{-1/2} \tag{22}$$

where d is a parameter (originally 4[86]). Note that for $A = B$, this reduces to eq 21. If $d = \infty$ and the geometric mean $\alpha_A\alpha_B$ is replaced by an arithmetic mean, this becomes the Ohno–Klopman formula that is widely used in semiempirical molecular orbital theory.[88–90]

Omitting the configuration interaction step, that is, using the lowest E_j rather than using eq 5, corresponds to assuming that the eigenvectors of the CI matrix are the configurations Φ_j and that the ground-state wave function is the CSF with the lowest energy. This procedure leads to a nondifferentiable potential energy surface whenever the identity of the lowest-energy CSF switches. The use of eq 5 eliminates this problem.

The total number of configurations of the form of eq 2 is factorially large and thus unmanageable. Therefore, we limit the sums in eq 5 to a subset of configurations. In the present work, we consider two options. In both cases, the reference CSF is the "aufbau" state where all electrons are paired up in the MOs with the lowest $E_{\mathrm{val}}^{(j)}$; systems with an odd number of electrons will have one MO with a single electron. The first set of excitations, whose use yields a method called TBCI-S, involves all single excitations from the occupied orbitals ($o_k^{(i)} = 1$ or 2) to the unoccupied or singly occupied orbitals ($o_k^{(j)} = 0$ or 1). In the second option, denoted TBCI-SPD, in addition to the single excitations, all paired double (PD) excitations are also included, where a paired double excitation is defined as an excitation of two paired electrons (i.e., electrons in the same orbital) from one doubly occupied orbital ($o_k^{(i)} = 2$) to the same unoccupied orbital ($o_k^{(j)} = 0$).

Finally, we consider $E^{(0)}$. The zero of energy is taken as the energy of the system with infinitely separated neutral atoms. Thus,

$$E^{(0)} = -NE_{\mathrm{atom}} \tag{23}$$

where N is the number of atoms and E_{atom} is the energy of a neutral atom. This same zero of energy is used regardless of the overall charge of the system.

The next issue is to approximate the CI coefficients in eqs 1, 4, and 5. We do this by using the following approximation:

$$|C_j|^2 = \frac{\exp\left(\frac{-E_j}{\Delta}\right)}{\sum_k \exp\left(\frac{-E_k}{\Delta}\right)} \tag{24}$$

where Δ is a parameter called the *resonance integral*. (Note that if one defines $\Delta = k_BT$, where k_B is the Boltzmann constant, then one would have the classic equation for a Boltzmann distribution.) These same weights are used in eq 4 to determine the partial atomic charges and in eq 5 to determine the overall energy. The role of Δ is to determine how much each CSF contributes to the total energy E. Those CSFs where $E_j - \min(E_j) \gg \Delta$ will have practically zero contributions, while those with $E_j - \min(E_j) \ll \Delta$ will have nearly equal contributions. Thus, one would like to have Δ on the same order as the spread of CSFs that one wants to contribute to the energy.

For simplicity, in the present work, we approximate χ_{ij} in eq 6 as a constant Δ, the same as the resonance integral in eq 24. Whether one might obtain better results by approximating χ_{ij} in terms of the overlap of configurations i and j is a subject for a future study. The roles of Δ and χ_{ij}, in their respective equations, are responsible (*vide infra*) for determining which CSFs contribute significantly. While the approximation that χ_{ij} is constant and is equal to Δ was initially made to reduce the parameter space during the parametrization, it also helps ensure size consistency.

Since both of the sums in eq 6 are over all excitations, the number of which scales as N^2, the computational effort of this method (eqs 5 and 6) scales as N^4. To reduce the scaling of the problem, only those terms j are included for which $|C_j|^2 \geq 10^{-8}$, as the rest contribute negligibly to the overall energy and its derivatives. For small systems, most or all of the excitations contribute; in this case, the computational effort for the method does scale as N^4, but this is not a concern due to the small sizes of the systems. This truncation, however, has a very dramatic effect on the scaling and computational times for larger systems, as most of the CSFs do not have a significant contribution to the overall energy. For example, for TBCI-S on Al_{177}, only 54 of the 117 838 CSFs (i.e., 0.05%) are considered. For TBCI-SPD, only 81 of the 235 410 (i.e., 0.03%) CSFs pass the cutoff criterion. In general, over 98% of the CSFs were found to be neglected. One thus has a very small prefactor on the N^4 step, with the rest of the work in eq 6 scaling as N^2, so that the computational cost for many systems will be dominated by the diagonalization of the tight-binding Hamiltonian matrix, an N^3-scaling process.

It was noted above that in eq 15 U_Z^l is taken as a parameter rather than as the experimental VSIP. This is motivated by considering the ionization potential of the Al atom, which should be just the VSIP of the Al 3p orbital. In TBCI, there is, however, the E_{CB} term (*vide supra*) that contains $\gamma_{AA}q_Aq_A$. Since $q_A = 1$ in the atomic cation and $q_A = 0$ in the neutral atom, this has a positive contribution to the cation and zero contribution to the neutral atom. In fact, one can show that for the Al atom:

$$U_{\mathrm{Al}}^s = -I_{\mathrm{Al}}^s + \frac{1}{\alpha_{\mathrm{Al}}} \tag{25}$$

$$U_{\mathrm{Al}}^p = -I_{\mathrm{Al}}^p + \frac{1}{\alpha_{\mathrm{Al}}} + 2\chi_{ij} \tag{26}$$

Equations 25 and 26 are used to determined the U_Z^l parameters in this work.

The method as described so far is suitable for only homonuclear systems, such as particles or clusters containing only aluminum atoms. When the theory is extended to heteronuclear systems, a number of entities become dependent on the atomic numbers of the atoms. Some of the variables—specifically U, α_A, and E_{atom}—depend only on a single atom's atomic number and become U_{Z_A}, α_{Z_A}, and $E_{\mathrm{atom}}(Z_A)$, respectively, where Z_A is the atomic number of atom A. Others depend on two atomic numbers, which may or may not be different, and become $\kappa_{Z_AZ_B}$, $\tau_{Z_AZ_B}$, $K_{Z_il_iZ_jl_jm}$, $\gamma_{Z_AZ_B}(R_{AB})$, and $d_{Z_AZ_B}$. Furthermore, the definition of the $E^{(0)}$ (eq 23) becomes

$$E^{(0)} = -\sum_Z N_Z\, E_{\mathrm{atom}}(Z) \tag{27}$$

where N_Z is the number of atoms of atomic number Z, and $E_{\mathrm{atom}}(Z)$ is the energy of a single atom. We will, however, continue in this article using the simpler notation that suffices for a homonuclear system.

One of the motivations behind developing TBCI was the problem, in TB, of improper charges when molecules dissociate. The fragment charges are obtained as the sum of the atomic charges of all of the atoms of a fragment. For the TBCI wave function, the atomic charges are obtained by combining eqs 4 and 18:

$$q_A = Z_A^* - \sum_{a\in A}\sum_j |C_j|^2 P_{aa}^{(j)} \tag{28}$$

where the weighting factors given by eq 24.

3. Analytical Gradients of TBCI

The derivative of the TBCI energy E (eq 5) with respect to a nuclear Cartesian coordinate X_C of atom C is

$$\frac{\partial E}{\partial X_C} = \sum_j \left(\frac{\partial |C_j|^2}{\partial X_C}\Gamma_j + |C_j|^2\frac{\partial \Gamma_j}{\partial X_C}\right) \tag{29}$$

From the equations for $|C_j|^2$ (eq 24) and Γ_j (eq 6), one finds that their derivatives are

$$\frac{\partial |C_j|^2}{\partial X_C} = \frac{|C_j|^2}{\Delta}\left(\frac{\sum_k\left[\exp\left(-\frac{E_k}{\Delta}\right)\frac{\partial E_k}{\partial X_C}\right]}{\sum_k \exp\left(-\frac{E_k}{\Delta}\right)} - \frac{\partial E_j}{\partial X_C}\right) \tag{30}$$

$$\frac{\partial \Gamma_j}{\partial X_C} = \frac{\partial E_j}{\partial X_C} + \frac{1}{2}\sum_i\left(\frac{\partial E_i}{\partial X_C} - \frac{\partial E_j}{\partial X_C}\right)\left(1 - \frac{E_i - E_j}{\sqrt{(E_i - E_j)^2 + 4\chi_{ij}^2}}\right) \tag{31}$$

From eq 12 for the energy of each CSF, one obtains an expression for $\partial E_j/\partial X_C$ that has three terms; since $E^{(0)}$ is independent of geometry, its gradients are zero. The gradients of the core−core repulsion energy (eq 14) are

$$\frac{\partial V_{\mathrm{rep}}}{\partial X_C} = -\sum_A\sum_{B>A} \tau V_{\mathrm{rep}}^{AB}\frac{\partial R_{AB}}{\partial X_C} \tag{32}$$

The gradients of the valence energy can be found using the Hellmann−Feynman theorem.[38] This theorem states that

$$\frac{\partial}{\partial X_C}\langle\Psi|\mathbf{H}|\Psi\rangle = \langle\Psi|\frac{\partial \mathbf{H}}{\partial X_C}|\Psi\rangle \tag{33}$$

with $\langle\Psi|\Psi\rangle = 1$.[91,92] Therefore, if

$$\mathbf{c}_k^{\dagger}\mathbf{H}\mathbf{c}_k = \varepsilon_k \tag{34}$$

where $\mathbf{c}_k$ is a column vector of the matrix with elements c_{ak}, then

$$\frac{\partial \varepsilon_k}{\partial X_C} = \mathbf{c}_k^{\dagger}\frac{\partial \mathbf{H}}{\partial X_C}\mathbf{c}_k \tag{35}$$

The one-electron Hamiltonian gradient matrix can be found by taking the derivatives of eqs 15−17, as in previous TB work.[93,94]

Finally, consider the gradients of the charge balance term. Applying the product rule to eq 19 yields

$$\frac{\partial E_{\mathrm{CI}}^{(j)}}{\partial X_C} = \sum_A \sum_{B \geq A}\left[\frac{\partial \gamma(R_{AB})}{\partial X_C} q_A^{(j)} q_B^{(j)} + \gamma(R_{AB})\,\frac{\partial q_A^{(j)}}{\partial X_C} q_B^{(j)} + \gamma(R_{AB})\, q_A^{(j)}\frac{\partial q_B^{(j)}}{\partial X_C}\right] \tag{36}$$

where the gradients of the Coulomb integrals $\gamma(R_{AB})$ can be found by taking the derivatives of eq 22:

$$\frac{\partial \gamma(R_{AB})}{\partial X_C} = e^2 R_{AB}\left[\alpha_A\alpha_B \exp\left(\frac{-R_{AB}^2}{d\alpha_A\alpha_B}\right) + R_{AB}^2\right]^{-3/2} \times \left(\frac{\exp\left(\frac{-R_{AB}^2}{d\alpha_A\alpha_B}\right)}{d} - 1\right)\frac{\partial R_{AB}}{\partial X_C} \tag{37}$$

Clearly, $\partial\gamma(R_{AB})/\partial X_C$ is nonzero only if X_C is a Cartesian coordinate of either atom A or B. The CSF atomic charges $q_A^{(j)}$ were found using a Mulliken−Coulson population analysis[80–82] (eq 18), and thus,

$$\frac{\partial q_A^{(j)}}{\partial X_C} = -\sum_{a \in A}\frac{\partial q_A^{(j)}}{\partial P_{aa}^{(j)}}\frac{\partial P_{aa}^{(j)}}{\partial X_C} \tag{38}$$

In the Supporting Information to a paper by Giesen et al.,[95] it is shown that

$$\frac{\partial q_A^{(j)}}{\partial P_{aa}^{(j)}} = -\sum_{b \in A}\delta_{ab} \tag{39}$$

From eq 11,

$$\frac{\partial P_{aa}^{(j)}}{\partial X_C} = \sum_k^{MO} 2\cdot o_k^{(j)} c_{ak}\frac{\partial c_{ak}}{\partial X_C} \tag{40}$$

The derivatives of the eigenvector matrix can be found using a unitary transformation as shown by Dykstra and Jasien:[96]

$$\mathbf{c}^{X_C} = \mathbf{c}\mathbf{U}^{X_C} \tag{41}$$

where $\mathbf{c}^{X_C}$ is the matrix of the derivatives of the c_{ak} coefficients with respect to the nuclear coordinate X_C, and $\mathbf{U}^{X_C}$ is the unitary transformation. Dykstra and Jasien showed that the off-diagonal elements of $\mathbf{U}^{X_C}$ are[96]

$$U_{ab}^{X_C}(\varepsilon_a - \varepsilon_b) = R_{ab}^{X_C}\varepsilon_b - G_{ab}^{X_C} \tag{42}$$

where

$$\mathbf{R}^{X_C} = \mathbf{c}^{\dagger}\mathbf{S}^{X_C}\mathbf{c} \tag{43}$$

$$\mathbf{G}^{X_C} = \mathbf{c}^{\dagger}\mathbf{H}^{X_C}\mathbf{c} \tag{44}$$

Since the overlap matrix is neglected in TBCI (i.e., $S_{ab} = \delta_{ab}$, *vide supra*), $\mathbf{R}^{X_C} = 0$. The diagonal elements of $\mathbf{U}^{X_C}$ are zero since[96]

$$U_{aa}^{X_C} = -\frac{1}{2}R_{aa}^{X_C} = 0 \tag{45}$$

4. The Aluminum Databases

Four databases were used in the parametrization and evaluation of the TBCI models. Each database has three components: an energy database, an IP database, and a cluster dissociation database. The complete databases are provided in the Supporting Information.

The largest database—Al974—is the union of several subsets. The first subset is called the Al808 database, and it consists of the 808 aluminum clusters and their energies given in a previously published database;[45] the composition of the 808 cluster database is given in Table 1. In addition, there are 22 ionization potentials of small Al_{1-13} clusters and of three larger clusters (Al_{19}, Al_{43}, and Al_{55}). There are also 34 dissociations of clusters ranging from Al_2 to Al_{16} and 15 additional dissociations involving larger clusters. The small neutral and cationic clusters that form the Al82 database (*vide infra*) are also included. This database (Al974) is used in the evaluation of the various theoretical methods.

The next database—Al824—is a subset of the larger database. It contains 686 of the 808 clusters, specifically Al_n

Table 1. Number of Clusters (n_k) of Each Size (N_k) in the Al808 Cluster Database

N_k	n_k	N_k	n_k	N_k	n_k	N_k	n_k	N_k	n_k
2	44	16	1	39	1	64	1	92	1
3	402	17	1	42	2	65	1	93	1
4	79	18	4	43	14	68	1	104	1
7	42	19	27	50	2	69	1	105	1
9	1	20	1	51	7	78	2	113	1
10	1	21	5	54	2	79	7	128	1
11	1	26	1	55	12	80	1	129	1
12	4	27	11	56	1	81	1	134	3
13	65	28	5	57	6	86	3	135	3
14	2	35	5	58	1	87	10	141	1
15	7	38	1	59	6	89	1	177	1

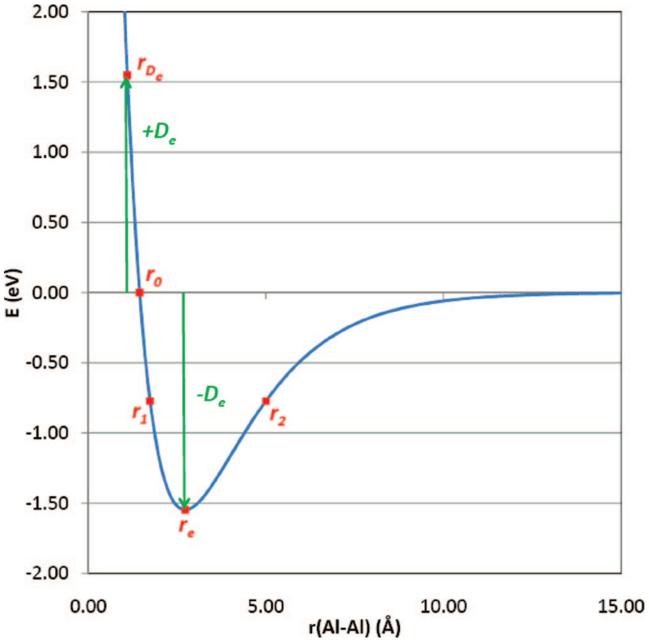


Figure 1. Diagram of a Morse curve showing r_e, r_{D_e}, r_0, r_1, and r_2.

where n = 2−13, 19, 35, 55, and 86. It also contains the Al82 database (*vide infra*). To this are added the 22 IPs and 34 dissociations involving small clusters only (*vide supra*). This database was used during the initial optimizations of the TBCI and TB+U methods.

The Al711 database, a subset of Al824, contains only the 711 clusters Al_n for which n = 2−13 (including those of appropriate size from the Al82 database). After significant preliminary work, it was found that this database is sufficient for optimizing the TBCI and TB+U methods (*vide infra*). This database does not include any IP or dissociation data.

Al82 is a database that was generated to fit the repulsive interaction of eq 14, as it is more balanced than the larger database for this purpose. For a number of reaction coordinates (stretch of Al_2, linear and perpendicular approach of Al to Al_2, symmetric stretch of D_{3h} Al_3, parallel and perpendicular approach of two Al_2 units, stretch of I_h Al_{13}, and an approach of two Al_9 units derived from a face-centered-cubic Al_{13} unit), five points (see Figure 1) were determined for both the neutral and cationic systems. First the minimum (r_e) of the potential energy surface was determined, along with the corresponding equilibrium dissociation energy (D_e). Then, using $V(r_e - 0.3$ Å) and $V(r = \infty)$, a Morse curve[97] was fit and the points $r_{D_e}(V = +D_e)$, r_0 ($V = 0$), and r_1 and r_2 ($V = -D_e/2$) were determined. The energies of all five points were determined at the PBEh/6-311+G(3d2f) or PBEh/MEC level of theory as appropriate (*vide supra*). For the smaller clusters ($n \leq 13$), this procedure was repeated for both the neutral and cationic systems; for the larger clusters, this procedure was done once for the neutral species, and the cationic species employ the same geometries as in the neutral case. Two extra points were added when, in two cases, the energy of the points r_{D_e} were too far from the predicted Morse curve.

The cluster energies and IPs were calculated at the PBEh/6-311+G(3d2f)[98−100] level of theory for clusters up to Al_{13} in size and at the PBEh/MEC level for larger clusters. The PBEh[101] hybrid exchange-correlation functional (Adamo and Barone's hybrid version[101] of the Perdew−Burke−Ernzerhof functional,[102,103] also called PBE0 or PBE1PBE) was chosen on the basis of comparisons[41] to accurate MCG3/3[104,105] energies for small clusters. The MEC basis set-relativistic effective core potential was designed to yield accurate energies for large aluminum clusters.[43]

The accuracy of the PBEh IP predictions was checked by comparing the IPs obtained for small clusters to IPs calculated using the coupled cluster *ab initio* method with all single and double substitutions[106] with a quasiperturbative estimate of the effect of the connected triple substitutions[107]—CCSD(T)—extrapolated to the complete-basis-set limit with the sequence of aug-cc-pV(n+d)Z basis sets. Dunning's aug-cc-pV(n+d)Z basis sets (n = D, T, Q)[108] were used as recommended by Martin et al. for the elements Al−Ar.[109,110] In particular, the Hartree−Fock (HF) energy, the CCSD correlation energy, and the connected triple excitations—(T)—contributions were extrapolated using the Weizmann-1 extrapolation scheme recently proposed by Martin and Parthiban:[111,112]

$$E_\infty = E_n + \frac{E_n - E_{n-1}}{(^{n}/_{n-1})^\beta - 1} \tag{46}$$

where the HF energy and CCSD contributions are determined with the two larger basis sets (i.e., $n = 4$) with β values of 5 and 3.22, respectively, while the (T) contribution is extrapolated with the two smaller basis sets (i.e., $n = 3$) and with $\beta = 3$. The comparison presented in Table 2 between the PBEh/6-311+G(3d2f) and extrapolated CCSD(T) IPs shows that the former are sufficiently accurate. The geometries of these specific Al clusters are given in the Supporting Information.

5. The Fitting Procedure

The error function for a given model and database has a number of components and is similar to that used in previous work.[39,41−43,45] The mean unsigned error per atom in the energies of a set of n_k aluminum clusters of size N_k is

$$\varepsilon_{N_k} = \frac{1}{2N_k}\left(\frac{\sum_{i=1}^{n_k} w_i \Delta E_i^k}{\sum_{i=1}^{n_k} w_i} + \frac{\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k} w_i w_j \Delta\Delta E_{ij}^k}{\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k} w_i w_j}\right) \tag{47}$$

where w_i is the weight of cluster i (*vide infra*) and

$$\Delta E_i^k = |E_i^{k,\mathrm{PBEh}} - E_i^{k,\mathrm{TBCI}}| \tag{48}$$

$$\Delta\Delta E_{ij}^k = |\Delta E_i^k - \Delta E_j^k| \tag{49}$$

and $E_i^{k,\mathrm{PBEh}}$ and $E_i^{k,\mathrm{TBCI}}$ are the PBEh and TBCI energies, respectively, of cluster i of size N_k. The second term in eq 47 was found in previous work[39,41,42,45] to be important in order to obtain a better fit with respect to the relative energies within a set of clusters of the same size. The total mean unsigned error in the energies is

$$\varepsilon_{\mathrm{EN}} = \frac{1}{N_{N_k}} \sum_{k=1}^{N_{N_k}} N_k \varepsilon_{N_k} \tag{50}$$

Table 2. Comparison of the PBEh/6-311+G(3d2f) and Extrapolated CCSD(T) Ionization Potentials (eV)

	extrap. CCSD(T)[a]	PBEh	ΔIP_i[b]	%Δ[c]
Al_2 (R_{Al-Al} = 1.9 Å)	7.701	7.761	0.060	0.8
Al_2 (R_{Al-Al} = 2.528202 Å)	6.570	6.600	0.030	0.5
Al_2 (R_{Al-Al} = 2.7 Å)	6.362	6.352	−0.010	−0.2
Al_3 $C_{\infty v}$ (R_{Al-Al} = 2.863, 1.699 Å)	6.068	6.053	−0.014	−0.2
Al_3 D_{3h} (R_{Al-Al} = 2.59 Å)	6.514	6.545	0.031	0.5
Al_3 D_{3h} (R_{Al-Al} = 2.5066 Å)	6.454	6.533	0.080	1.2
Al_4 (edge-on approach of Al (R_{Al-Al} = 2.629 Å) to Al_3 D_{3h} R_{Al-Al} = 2.5066 Å)	6.450	6.410	−0.040	−0.6
Al_4 (vertex approach of Al (R_{Al-Al} = 1.800 Å) to Al_3 D_{3h} R_{Al-Al} = 2.863 Å)	6.465	6.376	−0.089	−1.5
Al_4 (top-on approach of Al (R_{Al-Al} = 3.08 Å) to a slightly distorted Al_3 D_{3h} R_{Al-Al} = 2.5 Å)	6.038	6.098	0.060	1.0
Al_4 (rhomboid (D_{2d}) R_{Al-Al} = 2.551 Å)	6.690	6.560	−0.110	−1.6
Al_5 (C_{2v} planar)	6.616	6.576	−0.040	−0.6
MUE[d]		0.0513		

[a] Extrapolated CCSD(T); see text. [b] Difference in IPs; see eq 54. [c] Percent difference in IP. [d] Mean unsigned error in eV.

where N_{N_k} is the number of different cluster sizes. At certain times during the parametrization (*vide infra*), the contribution of $\varepsilon_{N_k=2}$ to ε_{EN} was increased by a factor of 5; when this is done, $\varepsilon_{N_k=2}$ is multiplied by 5, but N_{N_k} and n_k were not changed.

The databases contain structures that are high in energy due to small interatomic distances. To prevent these structures from dominating the fit, weights (w_i) are included in eq 47. A scheme previously used[45] in the parametrization of analytical potential energy functions for aluminum was chosen. This scheme is defined as follows. If R_i is the smallest interatomic distance in a given system, then w_i is defined as

$$w_i = \begin{cases} 1 & R_i \geq R_{\text{nc}} \\ \dfrac{V_2(R_{\text{nc}})}{V_2(R_i)} & R_i < R_{\text{nc}} \end{cases} \tag{51}$$

where R_{nc} is the nonclose radius, the cutoff for the definition of "too close," and $V_2(R)$ is the energy of the aluminum dimer at separation R. Rather than calculate, using DFT, the diatomic energy for every R_i, all of the available Al_2 DFT data, with $R \leq 3.0$ Å, were fit to a polynomial. The obtained polynomial, with a fitness of $R^2 = 0.998$, is

$$V_2(R_i) = -15.66R_i^3 + 114.0R_i^2 - 275.4R_i + 219.2 \tag{52}$$

where the units of V_2 and R_i are electronvolts and Ångstroms, respectively. R_{nc} is chosen such that $V_2(R_{\text{nc}}) = -V_2(r_e)$, which results in $R_{\text{nc}} = 1.798$ Å.

The MUE for the IPs is

$$\varepsilon_{\text{IP}} = \frac{1}{2}\left(\frac{1}{n_{\text{IP}}}\sum_{i=1}^{n_{\text{IP}}} \Delta\text{IP}_i + \frac{2}{n_{\text{IP}}(n_{\text{IP}}-1)} \sum_{i=1}^{n_{\text{IP}}-1} \sum_{j=i+1}^{n_{\text{IP}}} \Delta\Delta\text{IP}_{ij}\right) \tag{53}$$

where n_{IP} is the number of clusters in the IP database and

$$\Delta\text{IP}_i = |\text{IP}_i^{\text{PBEh}} - \text{IP}_i^{\text{TBCI}}| \tag{54}$$

$$\Delta\Delta\text{IP}_{ij} = |\Delta\text{IP}_i - \Delta\text{IP}_j| \tag{55}$$

The MUE of the fragment charges upon dissociation, ε_{dis}, is:

$$\varepsilon_{\text{dis}} = \frac{\Omega \dfrac{\text{hartree}}{\text{unit charge}}}{n_{\text{dis}}} \sum_{i=1}^{n_{\text{dis}}} |q_i| \tag{56}$$

where n_{dis} is the number of dissociating clusters in the database, q_i is the charge on each fragment of dissociated cluster i as determined by TBCI, and Ω is a constant taken as 0.1 on the basis of initial estimates of the relative magnitudes of ε_{EN}, ε_{IP}, and ε_{dis}. The errors in the dissociation charges are $|q_i|$ since experimentally only neutral fragments are observed. The total error function ε is defined as

$$\varepsilon = \frac{N_{N_k}\varepsilon_{\text{EN}} + \varepsilon_{\text{IP}} + \varepsilon_{\text{dis}}}{N_{N_k} + 1 + \Omega} \tag{57}$$

The error function in eq 47 is in units of electronvolts per atom, while all other error functions are in units of electronvolts.

In order to fit the parameters in the TBCI models, a microgenetic algorithm[113] was used, specifically version 1.7a of Carroll's FORTRAN code.[114] locally modified with our own fitness function and designed to run in parallel using the message-passing interface (MPI).[115,116] Because genetic algorithms, by definition, maximize a given function, the fitness function, f, used was minus the total error function ε.

There are different components to the TBCI model. Rather than optimize all of the parameters at once, it was decided to optimize the model in stages. Thus, for a given component being optimized, initial values for the other parameters are chosen on the basis of reasonable values—either from physically reasonable values or from a previous optimization—that were kept fixed during the optimization. The parameters in V_{rep} (i.e., κ and τ) were the first to be optimized. During the initial stages, values for the remaining parameters were chosen as follows:

- In E_{val}, the Wolfsberg−Helmholz parameters were taken from a previously published TB model,[46] specifically the third entry in Table S1 of this reference, which corresponds to a TB-WH model, where all of the Wolfsberg−Helmholz constants ($K_{l_i l_j m}$, eq 16) are given by a single K_0. The Slater-

Table 3. Final Parameters of the TBCI and TB+U Models

	model				model		
parameter (units)	S	SPD	TB+U	parameter (units)	S	SPD	TB+U
$K_{ss\sigma}$ (unitless)	0.033858	0.052569	0.045132	κ (eV)	796.05	573.02	564.15
$K_{sp\sigma}$ (unitless)	0.019876	0.37467	0.54078	τ (Å^{-1})	2.8216	2.6602	2.5734
$K_{pp\sigma}$ (unitless)	1.6238	1.8041	1.4896	α (Å)	3.7804	3.7804	3.7804
$K_{pp\pi}$ (unitless)	1.9347	1.2179	1.5931	d (unitless)	1.0	1.0	1.0
$\zeta_{ss\sigma}$ (Å^{-1})	2.5935	2.5935	2.5935	Δ (eV)	0.05	0.05	0.05
$\zeta_{sp\sigma}$ (Å^{-1})	2.5772	2.5772	2.5772	$E^{(0)}$ (eV)	−15.851	−15.951	−15.851
$\zeta_{pp\sigma}$ (Å^{-1})	2.5610	2.5610	2.5610	U_s (eV)	6.8110	6.8110	6.8110
$\zeta_{pp\pi}$ (Å^{-1})	2.5610	2.5610	2.5610	U_p (eV)	1.9770	1.7770	2.0770

type orbital exponents (ζ_a) were taken from the Hartree−Fock calculations by Clementi and Raimondi.[78] Thus, $K_0 = 0.409\,61$, $\zeta_s = 2.5935$ Å^{-1}, and $\zeta_p = 2.5610$ Å^{-1}.

• The diagonal Hamiltonian elements U_a (eq 15) were determined using eqs 25 and 26, where α_{Al} was determined as below, and the VSIPs were taken from experiments and are $I^s_{Al} = 10.620$ eV and $I^p_{Al} = 5.986$ eV;[117,118] thus, $U^s_{Al} = 6.811$ eV and $U^p_{Al} = 1.977$ eV.

• For γ_{AB} (eq 22), the three parameters were chosen on the basis of physical intuition; for simplicity, $d = 1.0$. Since α_{Al} is the distance where the repulsion switches from an r^{-1} to an e^{-r} behavior, and since this occurs as the two electron clouds start to overlap, α_{Al} was chosen as twice the van der Waals radius for Al, which has been experimentally determined to be 1.89 Å.[119]

• The parameter Δ controls how many CSFs contribute significantly to the configuration interaction wave function. It was chosen to be approximately 2 orders of magnitude smaller than the ionization potential of the aluminum atom; in particular, $\Delta = 0.10$ eV.

• The zero of energy, $E^{(0)}$, was determined from a calculation on the aluminum atom with the given set of parameters.

With the above parameters frozen, the two parameters in V_{rep} were optimized. For this optimization, the Al82 database was used and the quantity that was minimized was ε_{EN} (eq 50). Once these parameters were optimized, they were frozen and K_0 was then optimized. This was done using Al824 and fitting to ε (eq 57) rather than ε_{EN}. This cycle of optimizations was repeated. Finally, all three parameters in each model were allowed to vary. From these values, models were optimized with different values of Δ.

While the TBCI models based on TB-WH—especially TBCI-SPD—showed reasonable results, notably in the fragmentation charges, they are not sufficiently reliable. Therefore, the next level of TB approximation—EWH—was used for the TBCI and TB+U models. In this model, four different $K_{l,l',m}$ (eq 16) values are used. Because preliminary evaluations showed that the previous models were most deficient in the performance for small clusters, initially the TBCI models were optimized against ε_{EN} for the Al711 database. Finally, for both TBCI and TB+U, each with several fixed values of Δ, we simultaneously optimized six parameters (four Wolfsberg−Helmholz and two repulsion parameters) against ε for our Al824 database.

It was noted during the evaluation of the obtained models that, while the overall performance was satisfactory, the models predicted Al_2 to be too strongly bound by over half an electronvolt. By increasing the relative weight of ε_2 in ε_{EN} (see eq 50) 5-fold, models were obtained that showed improved performance for Al_2 without significantly compromising the fits to the rest of the data.

6. Computational Methods and Software

The CCSD(T) calculations were performed using the MOLPRO 2006.2 *ab initio* program package.[120] All DFT calculations were carried out using *Gaussian 03*,[121] except for the IP calculations for the clusters larger than Al_{55}, which were calculated using NWChem, version 4.5.[122] The TBCI calculations were done using an in-house code. The comparisons to the previously published NP-A and NP-B potentials[45] were done using published routines.[123] The comparisons to the TB and MBTB models[39,44] were done using the TB 2.0 code.[124]

The newly developed TBCI and TB+U models are implemented in TBPAC 2007,[125] which is available from the authors at http://comp.chem.umn.edu/tbpac/.

7. Results and Discussion

Three different TBCI models were examined: TBCI-S, TBCI-SPD, and TB+U; for TB+U, we used only single excitations. On the basis of extensive tests, we found that $\Delta = 0.05$ eV led to better results than $\Delta = 0.10$ eV and dramatically better results than $\Delta = 0.20$ eV; therefore, we chose $\Delta = 0.05$ eV for the final optimizations. We also set $\chi_{ij} = \Delta$. For all parameter sets in the present article, we also constrained $d = 1$ without optimization. All other parameters were optimized or frozen as discussed above. The final parameters are in Table 3.

In addition to comparisons between the TBCI models, the TBCI results are compared in Table 4 to three other kinds of results, with the comparison in all three cases based on comparing the errors measured against the PBEh Al974 database. The first and second kinds of methods to which we compare are the set of results obtained using TB. Previously, six different TB models were parametrized—three based on the Wolfsberg−Helmholz model in eqs 15 and 16 (TB-WH, TB-EWH, and TB-OWH) and three that contain many-body terms (MBTB, specifically TB-S, TB-CN, and TB-BA that include, respectively, screening, coordination number and bond angle many-body effects). Five parametrizations of the WH model were used, one from the original TB paper (herein denoted as TB-WH(SSST))[39] and four (denoted TB-WH(JST*i*), $i = 1-4$) that are the first four entries in Table S-1 of ref 46. The deficiencies of these TB

Table 4. Mean Errors[a]

model	ε_{EN} (0)	ε (0)	ε_{EN}	ε_{IP}	ε_{dis}	ε	max(q)[b]
TBCI-S	1.01	65.52	1.08	704.8	1.07	65.59	±6(×5)
TBCI-SPD	0.94	54.40	1.01	571.1	0.89	54.48	±5(×7)
TB+U[c]	0.99	86.71	1.12	1287.2	1.29	86.83	±6(×6)
TB-WH(SSST)	2.33	93.00	2.44	570.3	1.66	93.11	±6
TB-WH(JST1)	2.11	102.53	2.23	582.3	1.86[d]	102.65	±7(×4)[d]
TB-WH(JST2)	2.00	93.19	2.12	598.6	1.66[d]	93.30	±7(×2)[d]
TB-WH(JST3)	1.93	90.39	2.05	591.3	1.61[d]	90.50	±7[d]
TB-WH(JST4)	1.15	150.34	1.30	450.3	2.91	150.49	±10
TB-EWH	1.05	110.95	1.15	529.9	2.07	111.05	±8
TB-OWH	0.71	380.59	0.80	514.8	7.64	380.69	±36
MBTB-S	1.74	114.24	1.87	482.3	2.14	114.37	±9(×2)
MBTB-CN	2.01	357.36	2.09	403.8	7.18	357.45	±34
MBTB-BA	1.57	249.32	1.68	488.9	4.93	249.44	±18
NP-A[e]	0.71						
NP-B[e]	0.80						

[a] ε_{EN}, ε_{IP}, and ε are in units of meV. ε_{dis} is in units of charge. The first two columns with (0) exclude the data in Al82. [b] Maximum fragment charge of the clusters in the dissociation part of Al974. The number in parentheses denotes the number of instances of this maximum charge. [c] TB+U is with single excitations only. [d] All clusters for the methods in this row (in the case of TB-WH(JST3)-all but one) have at least ±1 charge. [e] Energies of neutral clusters only; see text.

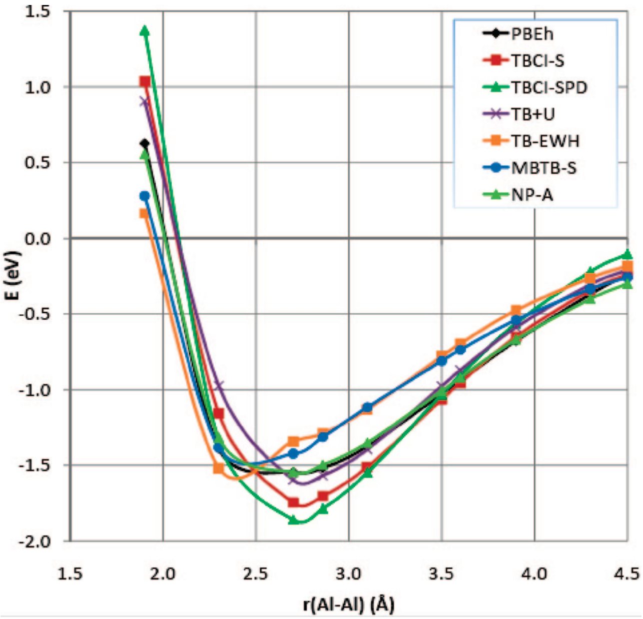


Figure 2. Plot of the Al_2 stretch potential energy surface for various theoretical methods.

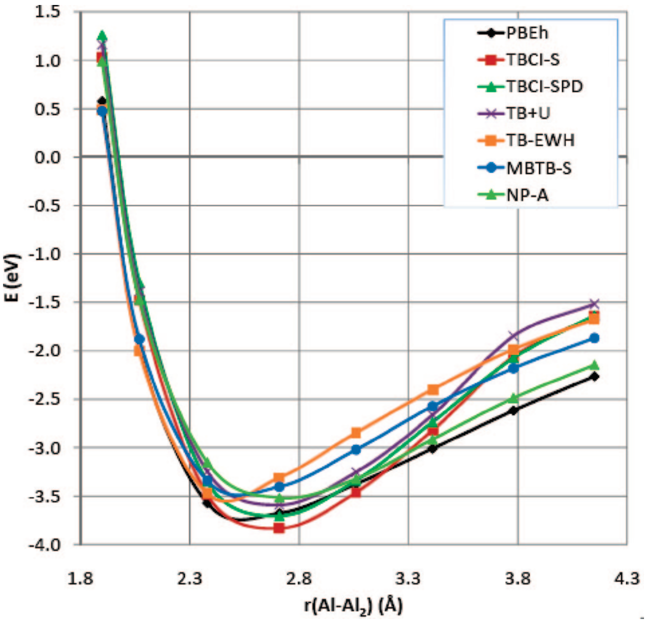


Figure 3. Plot of the perpendicular Al_2 + Al (C_{2v}, $r(Al_2)$ = 2.863 Å) potential energy surface for various theoretical methods.

methods, as well as the anticipated difficulty of extending them to heteronuclear systems, were the impetus behind this work; for a full description of these methods, see refs 39 and 46. The third kind of method to which we compare is the analytic PEFs, in particular NP-A and NP-B, which we previously developed.[45] These PEFs were found to accurately estimate the energy of aluminum clusters; since they do not contain any information on electrons or charges, they are incapable of predicting IPs or charges. NP-A is the more accurate of the two PEFs, and it includes many-body terms, while NP-B is an order of magnitude less computationally expensive but nearly as accurate.

Table 4 presents a comparison of the new and previous methods. This evaluation is over Al974. Also given in this table is an evaluation over Al974 with Al82 excluded. This was done because NP-A and NP-B cannot handle charged systems.

7.1. Preliminary Observations. Before discussing the final versions of the three new methods, we first mention some observations made during the parametrization.

First, we note that requiring all four Wolfsberg−Helmholz $K_{l_il_jm}$ constants to be the same (i.e., using the WH approximation, *vide supra*) leads to significantly larger (about a factor of 2 to 3) values of ε_{EN}, the average error in the energies of the clusters and nanoparticles. The optimized value of a single K is typically in the range of 0.55−0.8, but Table 3 shows that the final optimized values range from 0.03 to 1.93. In light of this wide range of optimized values, it is not surprising that a single compromise value is much worse. Removing the restriction that all four $K_{l_il_jm}$'s be equal (i.e., moving to the EWH approximation, *vide supra*) also lowered ε_{IP} by about a factor of 2.

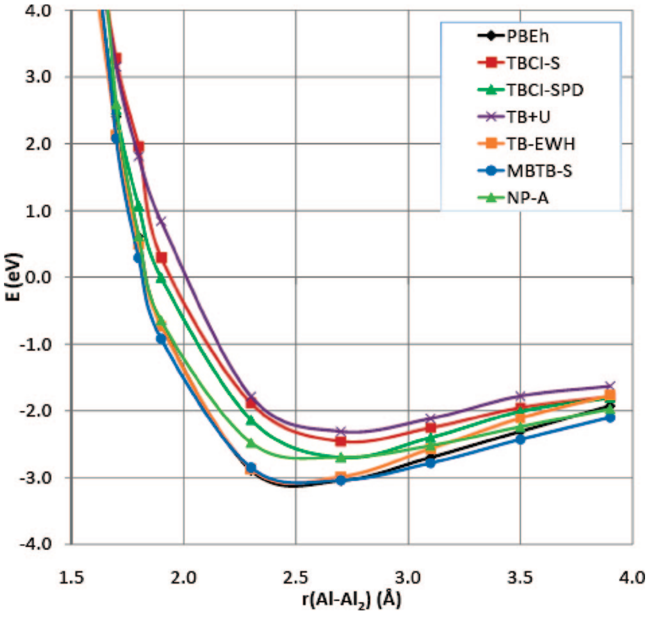


Figure 4. Plot of the linear Al + Al_2 ($C_{\infty v}$, $r(Al_2)$ = 2.863 Å) potential energy surface for various theoretical methods.

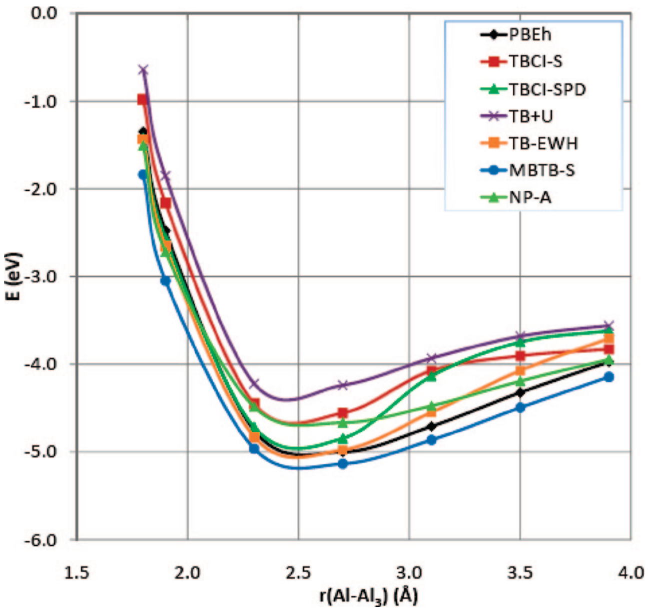


Figure 5. Plot of the on-top approach of Al to Al_3 (D_{3h}, r(Al−Al) = 2.863 Å) potential energy surface for various theoretical methods.

Another observation is that optimizing against only ε_{EN} for Al_n with $n = 2-13$ (i.e., Al711) is capable of yielding average errors (ε) for Al974 that are quite close (∼5−20% larger) to those obtained by optimizing over the larger Al824 database. This is an indication of the robustness of the methods, and thus the final optimizations were over this smaller subset.

It was also noted that, while the methods performed very well for the energies of Al clusters and nanoparticles, one glaring exception was Al_2. The initial parametrizations predicted Al_2 to be too strongly bound by more than 0.5 eV. A 5-fold increase of the relative weight of the Al_2 data in the evaluation of ε_{EN} (eq 50) provided TBCI and TB+U models that gave better predictions of the Al_2 potential energy surface with minimal (insignificant) deterioration of the remaining data predictions. A 10-fold increase provided excellent prediction of the Al_2 curve but resulted in poor potential energy curves for the other small clusters. Therefore, the final optimizations were against ε_{EN} for Al_{2-13} with a 5-fold increase in the weights for Al_2.

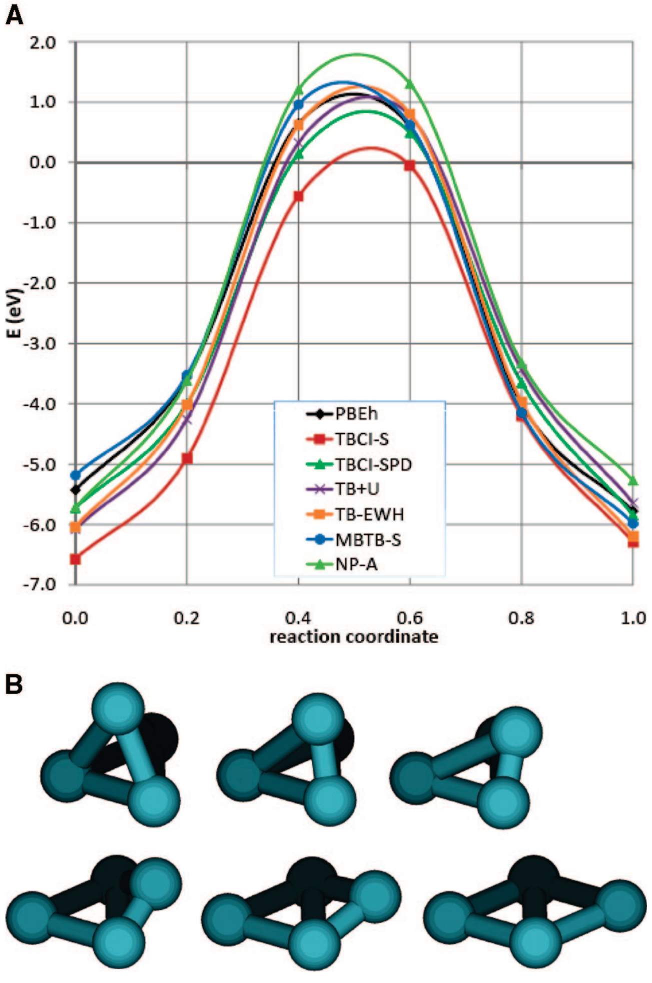


Figure 6. (A) Plot of the potential energy surface for the transition between tetrahedral (T_d) and rhomboid (D_{2h}) Al_4 for different theoretical methods. (B) Images of the clusters in the transition between tetrahedral and rhomboidal Al_4.

If we consider the TBCI-SPD calculations on all of the clusters in the energy and dissociation subsets of Al974, we find that the reference CSF (i.e, the aufbau CSF) is the dominant CSF in the CI expansion in only 57% of the cases.

7.2. Comparison of the Methods. Figures 2−7 depict potential energy profiles for various one-dimensional cuts through the potential energy surfaces of Al_2 to Al_7. These figures compare the results to PBEh, TB-EWH, MBTB-S, and NP-A. On the basis of these figures and Table 4, we can draw some conclusions.

First of all, we see that TBCI with the SPD configurational selection scheme is slightly better than the S scheme, on average, with the final ε decreasing by 17%. Both TBCI models yield similar errors in the cluster and nanoparticle energies (i.e., ε_{EN}), but TBCI-SPD is better in predicting IPs

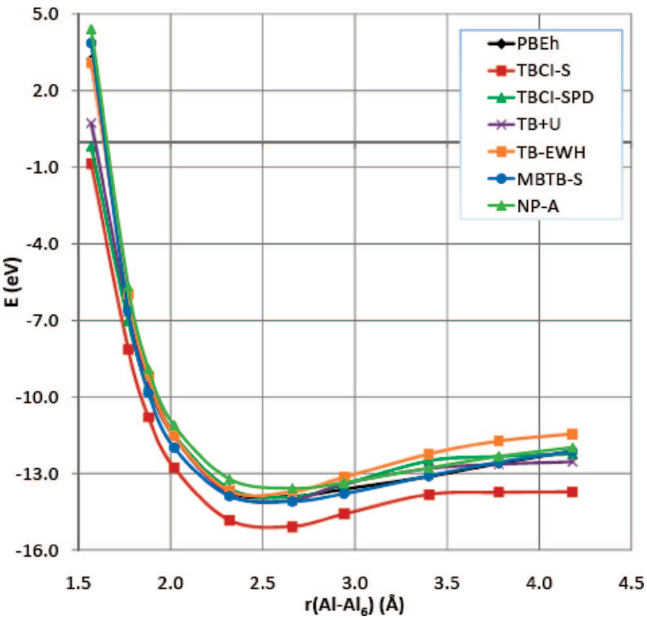


Figure 7. Plot of the edge approach of Al to Al_6 (O_h, r(Al−Al) = 2.8635 Å) potential energy surface for various theoretical methods.

(i.e., ε_{IP}) and fragment charges (i.e., ε_{dis}). TB+U, which is slightly faster than the TBCI models, performs equally well on predicting cluster and nanoparticle energies but is much worse in predicting fragmentation charges and is woeful in predicting IPs. In contrast to the traditional Hubbard model,[12] the TB+U method does not give the correct dissociation limit because the Hubbard (+U) correction is not introduced self-consistently in the present formalism.

In general, compared to the previously published TB and MBTB models,[39] the new methods are superior in predicting cluster and nanoparticle energies; the sole exception is the TB-OWH model. While TBCI-SPD and, to a lesser extent, TBCI-S are capable of predicting IPs with similar accuracy to that of the TB and MBTB methods, the new methods are superior when considering fragmentation charges. The TBCI and TB+U methods predict many of the cases to be neutral, while some of the TB methods are incapable of predicting any neutral fragments. In fact, TB-OWH predicts 15 of the dissociations to have fragment charges of ±10 or more (12 with ±20 or more).

It would be challenging to explain how TBCI is capable of eliminating such large charges. These charges are intrinsic artifacts of the methods, and not of our implementation, and affect also the older work of Slater and Koster,[1] Wolfsberg and Helmholz,[70] and Hoffmann.[2,72–75] In the TB models, the large charges result from the density of states in the valence energy region being more localized on one dissociation fragment. This would result in more electron density, and hence negative charge, being localized on this fragment, which in turn results in a positive charge localized on the other fragment. The excitations in TBCI could alleviate this charge imbalance by transferring electrons from one cluster to the other. However, only single and pair double excitations are considered, and thus, a maximum of two units of charge could be transferred. Nonetheless, much larger charges are being alleviated in the TBCI models. The most likely explanation is that the charge balance term prevents the Wolfsberg−Helmholz ($K_{l_il_jm}$) and repulsion (κ and τ) parameters from entering regions of parameter space, during their optimizations, that would result in large charge imbalances. Even more remarkable is the observation that reasonable fragmentation charges are obtained despite the fact that they were not included in the optimization of the TBCI models. In fact, when the TBCI models are evaluated against Al974 using the TB-OWH $K_{l_il_jm}$, $\zeta_{l_il_jm}$, and repulsion parameters, similarly large fragment charges are obtained as with TB-OWH. Thus, the charge balance term is clearly responsible for dampening, during the parametrization, any large charges that may occur. Note that, in this test of TB-OWH with CB terms, instead of eq 14, the following model for the repulsion was used here as in TB-OWH:

$$V_{\text{rep}} = \kappa \sum_{A} \sum_{B>A} \frac{\exp(-\tau R_{AB})}{R_{AB}^{u}} \tag{58}$$

where u is a constant.[39]

8. Summary

In summary, a new TB method has been proposed and developed. This model, called TBCI, improves on TB by applying a configuration-interaction-like procedure based on the TB orbitals. In such a manner, partial charges are incorporated into the calculation in a noniterative manner. This new TBCI model was optimized for aluminum nanoclusters and found to give exceptional performance with a low average error. The method is also applicable to other kinds of systems.

Acknowledgment. Research was supported in part by the Defense University Research Initiative on Nanotechnology (DURINT) through a grant managed by the Army Research Office. Computer resources were provided by the Minnesota Supercomputing Institute. M.A.I. also acknowledges support of a research grant from the Herbert J. Seligman Charitable Trust. A.H. gratefully acknowledges the Minnesota Supercomputing Institute for a research scholarship.

Supporting Information Available: The Al82 database, the ionization potentials and fragment dissociation components of Al824, the geometries of the clusters in Table 2, and the structures of the clusters in Figure 6. This material is available free of charge from the authors or via the Internet at http://pubs.acs.org.

References

(1) Slater, J. C.; Koster, G. F. *Phys. Rev.* **1954**, *94*, 1498–1524.

(2) Hoffmann, R. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

(3) *Tight-Binding Approach to Computational Materials Science*; Turchi, P. E. A., Gonis, A., Colombo, L., Eds.; MRS Symposium Proceedings 491, Materials Research Society: Warendale, PA, 1998.

(4) Harris, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31*, 1770–1779.

(5) Foulkes, W. M. C.; Haydock, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1989**, *39*, 12520–12546.

(6) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133−A1138.

(7) Hartree, D. R. *Proc. Cambridge Philos. Soc.* **1928**, *24*, 111–132.

(8) Slater, J. C. *Phys. Rev.* **1930**, *35*, 210–211.

(9) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

(10) Rein, R.; Fukuda, N.; Win, H.; Clarke, G. A. *J. Chem. Phys.* **1966**, *45*, 4743–4744.

(11) Newns, D. M. *Phys. Rev.* **1969**, *178*, 1123–1135.

(12) Hubbard, J. *Proc. R. Soc. London, Ser. A* **1963**, *276*, 238–257.

(13) Grimley, T. B.; Pisani, C. *J. Phys. C: Solid State Phys.* **1974**, *7*, 2831–2848.

(14) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenhein, T.; Suhai, S.; Seifert, G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 7260–7268.

(15) Frauenhein, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Status Solidi B* **2000**, *217*, 41–62.

(16) Micnas, R.; Ranninger, J.; Robaszkiewicz, S. *Rev. Mod. Phys.* **1990**, *62*, 113–171.

(17) Strack, R.; Vollhardt, D. *Phys. Rev. Lett.* **1993**, *70*, 2637–2640.

(18) Arrachea, L.; Aligia, A. A. *Phys. Rev. Lett.* **1994**, *73*, 2240–2243.

(19) Pulay, P. *Mol. Phys.* **1969**, *17*, 197–204.

(20) Fock, V. *Z. Phys.* **1930**, *61*, 126–148.

(21) Hartree, D. R.; Hartree, W. *Proc. R. Soc. London, Ser. A* **1935**, *150*, 9–33.

(22) Hartree, D. R.; Hartree, W. *Proc. R. Soc. London, Ser. A* **1936**, *154*, 588–607.

(23) Hartree, D. R. *Rep. Prog. Phys.* **1948**, *11*, 113–143.

(24) Hartree, D. R. *The Calculation of Atomic Structures*; John Wiley & Sons: New York, NY, 1957.

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

(26) Grimley, T. B. In *Adsorption at Solid Surfaces*; King, D. A., Woodruff, D. P., Eds.; Elsevier: Amsterdam, 1983; p 333.

(27) Shavitt, I. In *Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum Press: New York, NY, 1977; Vol. 4.

(28) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(29) Adams, G. F.; Bent, G. D.; Bartlett, R. J.; Purvis, G. D. In *Potential Energy Surfaces and Dynamics Calculations for Chemical Reactions and Molecular Energy Transfer*; Truhlar, D. G., Ed.; Plenum Press: New York, NY, 1981; pp 133−167.

(30) Head-Gordon, M. *J. Phys. Chem.* **1996**, *100*, 13213–13225.

(31) Shaik, S.; Hiberty, P. C. *Rev. Comput. Chem.* **2004**, *20*, 1–100.

(32) Truhlar, D. G. *J. Comput. Chem.* **2007**, *28*, 73–86.

(33) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718–8733.

(34) Brillouin, L. *Actual. Sci. Ind.* **1933**, *71*.

(35) Brillouin, L. *Actual. Sci. Ind.* **1934**, *159*.

(36) Nesbet, R. K. *Proc. R. Soc. London, Ser. A* **1955**, *230*, 312–321.

(37) Levy, B.; Berthier, G. *Int. J. Quantum Chem.* **1968**, *2*, 307–319.

(38) Simons, J.; Nichols, J. A. In *Quantum Mechanics in Chemistry*; Oxford University Press: New York, NY, 1997; p 442.

(39) Staszewska, G.; Staszewski, P.; Schultz, N. E.; Truhlar, D. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 045423.

(40) Valone, S. M.; Atlas, S. R. *J. Chem. Phys.* **2004**, *120*, 7262–7273.

(41) Schultz, N. E.; Staszewska, G.; Staszewski, P.; Truhlar, D. G. *J. Phys. Chem. B* **2004**, *108*, 4850–4861.

(42) Jasper, A. W.; Staszewski, P.; Staszewska, G.; Schultz, N. E.; Truhlar, D. G. *J. Phys. Chem. B* **2004**, *108*, 8996–9010.

(43) Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 41–53.

(44) Staszewska, G.; Staszewski, P.; Schultz, N. E.; Truhlar, D. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 039903(E).

(45) Jasper, A. W.; Schultz, N. E.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 3915–3920.

(46) Jasper, A. W.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 210–218.

(47) Li, Z. H.; Jasper, A. W.; Truhlar, D. G. *J. Am. Chem. Soc.* **2007**, *129*, 14899–14910.

(48) Li, Z. H.; Bhatt, D.; Schultz, N. E.; Siepmann, J. I.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, *111*, 16227–16242.

(49) Bhatt, D.; Schultz, N. E.; Jasper, A. W.; Siepmann, J. I.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 26135–26142.

(50) Bhatt, D.; Jasper, A. W.; Schultz, N. E.; Siepmann, J. I.; Truhlar, D. G. *J. Am. Chem. Soc.* **2006**, *128*, 4224–4225.

(51) Züttel, A.; Wenger, P.; Sudan, P.; Mauron, P.; Orimo, S.-i. *Mater. Sci. Eng., B* **2004**, *108*, 9–18.

(52) Ramaswamy, A. L.; Kaste, P.; Trevino, S. F. *J. Energ. Mater.* **2004**, *21*, 1–24.

(53) Ramaswamy, A. L.; Kaste, P. *J. Energ. Mater.* **2005**, *23*, 1–25.

(54) Galfetti, L.; De Luca, L. T.; Severini, F.; Meda, L.; Marra, G.; Marchetti, M.; Regi, M.; Bellucci, S. *J. Phys.: Condens. Matter* **2006**, *18*, S1991−S2005.

(55) Barnett, R. N.; Yannouleas, C.; Landman, U. *Z. Phys. D: At., Mol. Clusters* **1993**, *26*, 119–125.

(56) Mulder, F. M.; Thiel, R. C.; de Jongh, L. J.; Gubbens, P. C. M. *Nanostruct. Mater.* **1996**, *7*, 269–292.

(57) Kara, A.; Rahman, T. S. *Phys. Rev. Lett.* **1998**, *81*, 1453–1456.

(58) Link, S.; El-Sayed, M. A. *Int. Rev. Phys. Chem.* **2000**, *19*, 409–453.

(59) Voisin, C.; Del Fatti, N.; Christofilos, D.; Vallée, F. *J. Phys. Chem. B* **2001**, *105*, 2264–2280.

(60) *Metal Nanoparticles: Synthesis, Characterization, and Applications*; Feldheim, D. L., Foss, C. A., Jr., Eds.; Marcel Dekker: New York, NY, 2002.

(61) *Nanopartilces: From Theory to Application*; Schmid, G., Ed.; Wiley-VCH: Weinheim, Germany, 2004.

(62) Mitev, P.; Papageorgiou, D. G.; Lekka, C. E.; Evangelakis, G. A. *Surf. Sci.* **2004**, *566−568*, 937–943.

(63) Mazzone, A. M.; Morandi, V. *Comput. Mater. Sci.* **2007**, *38*, 830–837.

(64) Sun, C. Q. *Prog. Solid State Chem.* **2007**, *35*, 1–159.

(65) Jiang, Q.; Ao, Z. M.; Zheng, W. T. *Chem. Phys. Lett.* **2007**, *439*, 102–104.

(66) Assfeld, X.; Almlöf, J. E.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *241*, 438–444.

(67) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S136−S151.

(68) Klopman, G.; Evans, R. C. In *Semiempirical Methods of Electronic Structure Theory, Part A*; Segal, G. A.; Plenum Press: New York, NY, 1977; pp 29−67.

(69) Wang, Y.; Mak, C. H. *Chem. Phys. Lett.* **1995**, *235*, 37–46.

(70) Wolfsberg, M.; Helmholz, L. *J. Chem. Phys.* **1952**, *20*, 837–843.

(71) Mulliken, R. S. *J. Chim. Phys. Phys. Chim. Biol.* **1949**, *46*, 497–542.

(72) Hoffmann, R. *J. Chem. Phys.* **1964**, *40*, 2047–2048.

(73) Hoffmann, R. *J. Chem. Phys.* **1964**, *40*, 2474–2480.

(74) Hoffmann, R. *J. Chem. Phys.* **1964**, *40*, 2480–2488.

(75) Hoffmann, R. *J. Chem. Phys.* **1964**, *40*, 2745–2745.

(76) Slater, J. C. *Phys. Rev.* **1930**, *36*, 57–64.

(77) Jones, H. W. *Int. J. Quantum Chem.* **1981**, *19*, 567–574.

(78) Clementi, E.; Raimondi, D. L. *J. Chem. Phys.* **1963**, *38*, 2686–2689.

(79) Jones, H. W. *Int. J. Quantum Chem.* **1980**, *18*, 709–713.

(80) Mulliken, R. S. *J. Chem. Phys.* **1935**, *3*, 564–573.

(81) Coulson, C. A.; Longuet-Higgins, H. C. *Proc. R. Soc. London, Ser. A* **1947**, *191*, 39–60.

(82) Maslen, V. W.; Coulson, C. A. *J. Chem. Soc.* **1957**, 4041–4049.

(83) Liechtenstein, A. I.; Anisimov, V. I.; Zaanen, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1995**, *52*, R5467–R5470.

(84) Leung, K.; Rempe, S. B.; Schultz, P. A.; Sproviero, E. M.; Batista, V. S.; Chandross, M. E.; Medorth, C. J. *J. Am. Chem. Soc.* **2006**, *128*, 3659–3668.

(85) Pariser, R.; Parr, R. G. *J. Chem. Phys.* **1953**, *21*, 767–776.

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

(87) Cramer, C. J.; Truhlar, D. G. *Rev. Comput. Chem.* **1995**, *6*, 1–72.

(88) Ohno, K. *Theor. Chim. Acta* **1964**, *2*, 219–227.

(89) Klopman, G. *J. Am. Chem. Soc.* **1964**, *86*, 4550–4557.

(90) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164–170.

(91) Hellman, J. *Einführung in die Quantenchemie*; Deuticke & Co.: Leipzig, 1937.

(92) Feynman, R. P. *Phys. Rev.* **1939**, *56*, 340–343.

(93) Liu, T. A Tight Binding Model for the Energetics of Hydrocarbon Fragments on Metal Surfaces. Ph.D. Thesis, University of Minnesota, Minneapolis, MN, May 2000.

(94) Liu, T.; Truhlar, D. G. Unpublished results.

(95) Giesen, D. J.; Storer, J. W.; Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1995**, *117*, 1057–1068.

(96) Dykstra, C. E.; Jasien, P. G. *Chem. Phys. Lett.* **1984**, *109*, 388–393.

(97) Morse, P. M. *Phys. Rev.* **1929**, *34*, 57–64.

(98) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294–301.

(99) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.

(100) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639–5648.

(101) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

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

(103) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

(104) Fast, P. L.; Sánchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407–410.

(105) Tratz, C. M.; Fast, P. L.; Truhlar, D. G. *PhysChemComm* **1999**, *2*, 70–79.

(106) Purvis, G. D., III; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

(107) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(108) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.

(109) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129–4141.

(110) Martin, J. M. L. *THEOCHEM* **2006**, *771*, 19–26.

(111) Martin, J. M. L.; de Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843–1856.

(112) Martin, J. M. L.; Parthiban, S. In *Quantum Mechanical Prediction of Thermochemical Data*; Cioslowski, J., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001; pp 31−65.

(113) Carroll, D. L. In *Developments in Theoretical and Applied Mechanics*; Wilson, H., Batara, R., Bert, C., Davis, A., Schapery, R., Stewart, D., Swinson, F., Eds.; School of Engineering, The University of Alabama: Tuscaloosa, AL, 1996; Vol *XVII*, p 411.

(114) Carroll, D. L. GA_version 1.7a: FORTRAN Genetic Algorith, Driver; CU Aerospace: Urbana, IL, 2001. See: http://cuaerosopace.com/carroll/ga.html (accessed Mar 2008).

(115) Message Passing Interface Forum. MPI: A message-passing interface standard. Computer Science Department Technical Report CS-93-214; University of Tennessee, Knoxville, TN, November 1993.

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

(117) *CRC Handbook of Chemistry and Physics*, 78th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 1997.

(118) Moore, C. E. *Atomic Energy Levels as Derived from the Analyses of Optical Spectra, Circular of the National Bureau of Standards 467*; U.S. Department of Commerce: Washington, DC, 1949; Vol. 1.

(119) Behm, J. M.; Blume, T.; Morse, M. D. *J. Chem. Phys.* **1994**, *101*, 5454–5463.

(120) Werner, H. J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. MOLPRO, version 2006.2, a package of ab initio programs. See: http://www.molpro.net (accessed Mar 2008).

(121) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.01; Gaussian, Inc.: Pittsburgh, PA, 2003.

(122) Straatsma, T. P.; Apra, E.; Windus, T. L.; Dupuis, M.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Smith, D. M. A.; Hackler, M. T.; Pollack, L.; Harrison, R. J.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Valiev, M.; 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.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, version 4.5; Pacific Northwest National Laboratory: Richland, WA, 2003.

(123) Duchovic, R. J.; Volobuev, Y. L.; Lynch, G. C.; Jasper, A. W.; Truhlar, D. G.; Allison, T. C.; Wagner, A. F.; Garrett, B. C.; Espinosa-García, J.; Corchado, J. C. POTLIB-online. http://comp.chem.umn.edu/potlib (accessed Mar 2008).

(124) Staszewska, G.; Liu, T.; Truhlar, D. G. *TB*, version 2.0; University of Minnesota: Minneapolis, MN, 2003.

(125) Iron, M. A.; Staszewska, G.; Liu, T.; Jasper, A. W.; Truhlar, D. G. *TBPAC 2007*; University of Minnesota: Minneapolis, MN, 2007.

CT700343T

JCTC Journal of Chemical Theory and Computation

The MARTINI Coarse-Grained Force Field: Extension to Proteins

Luca Monticelli,[†] Senthil K. Kandasamy,[‡] Xavier Periole,[§] Ronald G. Larson,[‡]
D. Peter Tieleman,[†] and Siewert-Jan Marrink*,[§]

Dept of Biological Sciences, University of Calgary, 2500 University Dr NW, Calgary, AB, T2N 1N4, Canada, Chemical Engineering Department, The University of Michigan, 2300 Hayward Street, Ann Arbor, Michigan 48109, and Groningen Biomolecular Sciences and Biotechnology Institute & Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

Received November 27, 2007

Abstract: Many biologically interesting phenomena occur on a time scale that is too long to be studied by atomistic simulations. These phenomena include the dynamics of large proteins and self-assembly of biological materials. Coarse-grained (CG) molecular modeling allows computer simulations to be run on length and time scales that are 2–3 orders of magnitude larger compared to atomistic simulations, providing a bridge between the atomistic and the mesoscopic scale. We developed a new CG model for proteins as an extension of the MARTINI force field. Here, we validate the model for its use in peptide-bilayer systems. In order to validate the model, we calculated the potential of mean force for each amino acid as a function of its distance from the center of a dioleoylphosphatidylcholine (DOPC) lipid bilayer. We then compared amino acid association constants, the partitioning of a series of model pentapeptides, the partitioning and orientation of WALP23 in DOPC lipid bilayers and a series of KALP peptides in dimyristoylphosphatidylcholine and dipalmitoylphosphatidylcholine (DPPC) bilayers. A comparison with results obtained from atomistic models shows good agreement in all of the tests performed. We also performed a systematic investigation of the partitioning of five series of polyalanine−leucine peptides (with different lengths and compositions) in DPPC bilayers. As expected, the fraction of peptides partitioned at the interface increased with decreasing peptide length and decreasing leucine content, demonstrating that the CG model is capable of discriminating partitioning behavior arising from subtle differences in the amino acid composition. Finally, we simulated the concentration-dependent formation of transmembrane pores by magainin, an antimicrobial peptide. In line with atomistic simulation studies, disordered toroidal pores are formed. In conclusion, the model is computationally efficient and effectively reproduces peptide−lipid interactions and the partitioning of amino acids and peptides in lipid bilayers.

1. Introduction

Molecular simulations are a useful tool in the interpretation of experimental data, and they provide structural and dynamic details that cannot be easily probed experimentally. Despite the progress in computer hardware and simulation algorithms, atomistic simulations are still limited to systems containing tens or hundreds of thousands of atoms and a submicrosecond time scale. Cellular processes, however, cover time scales of nanoseconds to seconds and involve hundreds of different molecules interacting on a multitude of length scales. Many

* Corresponding author e-mail: s.j.marrink@rug.nl.
[†] University of Calgary.
[‡] The University of Michigan.
[§] University of Groningen.

10.1021/ct700324x CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/16/2008

biologically interesting phenomena, including vesicle fusion, formation of higher-order protein complexes, protein folding, and signal transduction, are beyond the capabilities of atomistic simulations.[1] In order to simulate these motions, simplification of the model is required. The use of coarse-grained (CG) models represents an attractive alternative to atomistic models, allowing for simulations to be run on larger systems and longer time scales and still providing some realistic structural details.

A large diversity of coarse-graining approaches for biomolecular systems is available. They range from qualitative, solvent-free models to models including chemical specificity. With reference to proteins, numerous coarse-grained descriptions have been developed for studying protein-folding thermodynamics and kinetics and protein structure prediction. These approaches include both structure-based, knowledge-based, and physics-based models. The type of approach and the level of sophistication of the different models vary greatly depending on the scope of the model and the properties investigated. Among structure-based models, simple Go models[2] proved useful in the characterization of the kinetics and cooperativity effects in protein folding. Miyazawa and Jernigan developed a knowledge-based statistical potential to predict the structure of proteins in solution.[3] This statistical potential was derived by estimating effective inter-residue contact energies from the numbers of residue–residue contacts observed in crystal structures of globular proteins. Das et al. recently developed a sophisticated knowledge-based potential incorporating sequence details and energetic frustration for a more realistic study of folding pathways.[4] Elastic network models[5,6] have been used in conjunction with normal-mode analysis to predict large-scale motions in proteins. They rely on the knowledge of the protein structure and on the assumption that motions relevant for biological functions depend on low-frequency collective fluctuations. Simple, exact physics-based models for proteins were pioneered by Chan and Dill.[7,8] Their lattice model was used to show that the basic features determining a protein's fold lie mainly in the topological arrangement of hydrophobic and polar residues. The use of physics-based coarse-grained models for protein structure prediction was pioneered by Warshel and Levitt[9] and later developed by many others. Among them, Scheraga and co-workers have developed a united-residue force field parametrized against restricted free-energy functions from all-atom simulations of polypeptide chains, without any information from structural databases.[10] Very different approaches have been used to study problems in which the protein structure does not need to be predicted and the details of it are not crucial. Coarse-grained models have been developed recently in the group of McCammon to study large-scale protein motions in HIV-1 protease.[11] Schulten and co-workers developed a mesoscopic protein model aimed at simulating large-scale motions of macromolecular assemblies. In this case, the protein is simply considered as an elastic object with a well-defined three-dimensional shape, and changes in the detailed protein structure are not accounted for.[12] A number of mesoscopic models have been developed in order to study the effect of membrane proteins on the properties of biological membranes, reviewed in refs13–16. These models generally disregarded the details of the protein structure and internal dynamics but proved useful in understanding protein–lipid interactions and lipid-mediated protein–protein interactions, progressing beyond the original lattice models. Both Venturoli et al. and Smeijers et al. developed a coarse-grained model that enabled the investigation of the effects of a mismatch in the hydrophobic thickness of proteins and the lipid bilayer (the so-called hydrophobic mismatch) and, in particular, the lipid-induced protein tilt and protein-induced membrane deformations.[17,18] Multiscale approaches that couple the atomistic and coarse-grained levels of description have also been applied recently to study peptide−membrane interactions[19] and proteins.[20,21] When all of the different coarse-grained approaches are considered, the level of resolution and the degree of flexibility of the proteins vary from several particles per amino acid to one particle per protein, depending on the size of the object and motions that need to be described.

Marrink and co-workers recently developed a coarse-grained force field for simulation of lipids and surfactants,[22,23] coined the MARTINI force field. The force field has been shown to reproduce semiquantitatively fundamental structural and thermodynamic properties of lipid bilayers.[22,24–26] The model was developed in close connection with atomistic models, but with a very different philosophy compared to iterative Boltzmann inversion,[27] inverted Monte-Carlo schemes,[28–30] or force matching[31] approaches. Instead of focusing on an accurate reproduction of structural details of a particular state for a specific system, the model aimed for a broader range of applications without the need to reparameterize the model each time. This was achieved by extensive calibration of the chemical building blocks of the coarse-grained force field against thermodynamic data, in particular, oil/water partitioning coefficients.[22,23] This is similar in spirit to the recent development of the GROMOS force field.[32]

Here, we present an extension of the MARTINI force field,[23] to model proteins. The overall aim of our coarse-graining approach is to provide a simple model that is computationally fast and easy to use, yet flexible enough to be applicable to a large range of biomolecular systems. In the MARTINI model, several atoms are grouped together in a "virtual" bead that interacts through an effective potential. The reduction of the number of degrees of freedom and the use of shorter-range potential functions makes the model computationally very efficient, allowing for a reduction of the simulation time by 2∼3 orders of magnitude compared to the most common atomistic models. The present model for proteins was developed using the same philosophy as for the lipids, using the partitioning free energy of amino acid side chains between water and oil phases to select the appropriate nonbonded interaction parameters. Processes such as protein folding, peptide membrane binding, and protein–protein recognition depend critically on the degree to which the constituents partition between polar and nonpolar environments. The choice of particle types and the nonbonded interaction matrix is left unaltered, making the protein force field fully compatible with the lipid force field.

The choice of the bonded parameters was based on the distribution of bond lengths, angles, and dihedrals calculated from the Protein Data Bank (PDB). Similar models have been developed recently by other groups as well. Starting from Marrink's original model for lipids,[22] the groups of Schulten and Sansom built a model for proteins and studied lipoprotein particles[33] and membrane proteins.[34,35] The major differences with the approach presented in this paper is that (*i*) we base our CG protein model on the new MARTINI CG force field, which has many more particle types and allows for discrimination between all amino acids, and (*ii*) we base the particle assignment on a systematic investigation of thermodynamic properties of each amino acid. Using a preliminary version of the current model, Periole et al.[36] recently managed to study the oligomerization of rhodopsins, a transmembrane protein belonging to the class of G-protein coupled receptors. It was found that the presence of hydrophobic mismatch favors rhodopsin aggregation, in quantitative agreement with results from FRET experiments that were performed in conjunction to the simulations. The simulations furthermore revealed that protein–protein interactions inside a membrane bilayer show a site preference related to localized mismatch, pointing to the importance of modeling proteins as chemically detailed objects rather than as simplified rods. Yefimov et al.[37] succeeded in simulating the spontaneous tension-driven gating of a membrane-embedded mechanosensitive protein channel, also using a prerelease of the MARTINI protein force field. This simulation comprises one of the first examples of a membrane protein in action, resolved at near-atomic detail. In another recent application,[38] the gating motions of membrane-embedded potassium channels were studied. It was found that channel gating is coupled to subtle displacements of the voltage sensor domain. A preliminary version of the current force field has also been applied to study the conformation of apoA-1 in model spheroidal high-density lipoprotein particles.[39] Extensive comparison of the CG system to all-atom simulations revealed a close correspondence, both in structure and in dynamics.

The present work is organized as follows. First, we describe the force field parametrization procedure, for both bonded and nonbonded interactions. Then, we present results for a range of test cases of the model, focusing on simulations of peptide-membrane systems. In particular, we show that (*i*) the potential of mean force for single amino acid side chains across a lipid membrane is very similar with the CG model compared to results obtained with all-atom models, (*ii*) the correct partitioning and orientation of a large variety of small peptides at the water−bilayer interface is reproduced, and (*iii*) antimicrobial peptides (AMPs) form transmembrane pores that look similar to what has been shown with all-atom simulations.

2. The Model

2.1. Basic Parametrization. The basic parameters for the CG peptide model are the same as those published previously for the CG lipid model.[22,23] The peptide force field described here is fully compatible with the latest lipid force field, coined the MARTINI force field. The version described in the current paper is denoted v2.1. In this section, we provide a brief overview of the basic parametrization. More details about the CG model can be found in the original papers.[22,23]

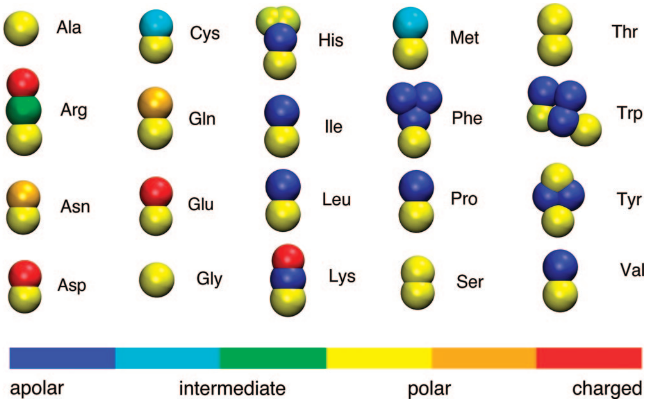


Figure 1. Coarse-grained representation of all amino acids. Different colors represent different particle types.

The Mapping. The MARTINI model[23] is based on a four-to-one mapping; that is, on average, four heavy atoms are represented by a single interaction center, with an exception for ringlike molecules. To map the geometric specificity of small ringlike fragments or molecules (e.g., benzene, cholesterol, and several of the amino acids), the general four-to-one mapping rule is insufficient. Ringlike molecules are therefore mapped with higher resolution (up to two-to-one). The model considers four main types of interaction sites: polar (P), nonpolar (N), apolar (C), and charged (Q). Within a main type, subtypes are distinguished either by a letter denoting the hydrogen-bonding capabilities (*d* = donor, *a* = acceptor, *da* = both, *0* = none) or by a number indicating the degree of polarity (from 1 = lower polarity to 5 = higher polarity). The mapping of all protein amino acids is shown in Figure 1.

Nonbonded Interactions. All particle pairs *i* and *j* at distance r_{ij} interact via a Lennard-Jones (LJ) potential:

$$V_{\text{Lennard-Jones}}(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1}$$

The strength of the interaction, determined by the value of the well depth ε_{ij} depends on the interacting particle types. The value of ε ranges from $\varepsilon_{ij} = 5.6$ kJ/mol for interactions between strongly polar groups to $\varepsilon_{ij} = 2.0$ kJ/mol for interactions between polar and apolar groups mimicking the hydrophobic effect. The effective size of the particles is governed by the LJ parameter: $\sigma = 0.47$ nm for all normal particle types. For the special class of particles used for ringlike molecules, slightly reduced parameters are defined to model ring–ring interactions: $\sigma = 0.43$ nm and ε_{ij} is scaled to 75% of the standard value. The full interaction matrix can be found in the original publication.[23] In addition to the LJ interaction, charged groups (type Q) bearing a charge *q* interact via a Coulombic energy function with a relative dielectric constant $\varepsilon_{\text{rel}} = 15$ for explicit screening:

$$V_{\text{el}} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_{\text{rel}} r_{ij}} \tag{2}$$

To avoid generation of unwanted noise, the nonbonded potential energy functions are used in their shifted form, in which both the energy and force vanish at the cutoff distance $r_{\mathrm{cut}} = 1.2$ nm. The LJ potential is shifted from $r_{\mathrm{shift}} = 0.9$ nm to r_{cut}. The electrostatic potential is shifted from $r_{\mathrm{shift}} = 0.0$ nm to r_{cut}. Shifting of the electrostatic potential in this manner mimics the effect of a distance-dependent screening. Nonbonded interactions between nearest neighbors are excluded.

Bonded Interactions. Bonded interactions are described by the following set of potential energy functions acting between bonded sites i, j, k, and l with equilibrium distance d_{b}, angle ϕ_{a}, and dihedral angles ψ_{d} and ψ_{id}:

$$V_{\mathrm{b}} = \frac{1}{2}K_{\mathrm{b}}(d_{ij} - d_{\mathrm{b}})^2 \qquad (3)$$

$$V_{\mathrm{a}} = \frac{1}{2}K_{\mathrm{a}}[\cos(\varphi_{ijk}) - \cos(\varphi_{\mathrm{a}})]^2 \qquad (4)$$

$$V_{\mathrm{d}} = K_{\mathrm{d}}[1 + \cos(n\psi_{ijkl} - \psi_{\mathrm{d}})] \qquad (5)$$

$$V_{\mathrm{id}} = K_{\mathrm{id}}(\psi_{ijkl} - \psi_{\mathrm{id}})^2 \qquad (6)$$

The force constants K are generally weak, inducing flexibility of the molecule at the coarse-grained level mimicking the collective motions at the fine-grained level. The bonded potential V_{b} is used for chemically bonded sites and the angle potential V_{a} to represent chain stiffness. The improper dihedral angle potential V_{id} is used to prevent out-of-plane distortions of planar groups. Proper dihedrals V_{d} are used to impose secondary structure of the peptide backbone. It is important to note, therefore, that in the current parametrization conformational changes of protein secondary structure are not adequately modeled.

2.2. Mapping of the Amino Acids. The mapping of all 20 amino acids is depicted in Figure 1 and presented in Table 1. Most amino acids are mapped onto single standard particle types in a similar way as was done recently by other groups.[33,34] The apolar amino acids (Leu, Pro, Ile, Val, Cys, and Met) are represented as C-type particles, the polar uncharged amino acids (Thr, Ser, Asn, and Gln) by P-type particles, and the amino acids with small negatively charged side chains as Q-type (Glu and Asp). The positively charged amino acids Arg and Lys are modeled by a combination of a Q-type particle and an uncharged particle. The bulkier ring-based side chains are modeled by three (His, Phe, and Tyr) or four (Trp) beads of the special class of ring particles. The Gly and Ala residues are only represented by the backbone particle. The type of the backbone particle depends on the protein secondary structure (see Table 2); free in solution or in a coil or bend, the backbone has a strong polar character (P type); as part of a helix or β strand, the interbackbone hydrogen bonds reduce the polar character significantly (N type). Proline is less polar due to the lack of hydrogen-donor capabilities.

Table 1. Mapping of the Amino Acids and Free Energy of Partitioning between Water and Butane (Calculated) or Water and Cyclohexane (Experimental Measure[40,41])

			free energy (kJ/mol)	
side chain	CG representation	mapping scheme[a]	CG	exptl.
Leu	C1[b]		22	22
Ile	C1[b]		22	22
Val	C2[b]		20	17
Pro	C2[b]		20	
Met	C5		9	10
Cys	C5		9	5
Ser	P1		−11	−14
Thr	P1		−11	−11
Asn	P5		< −25	−28
Gln	P4		−23	−25
Asp	Qa		< −25	
Asp (uncharged)	P3		−18	−19
Glu	Qa		< −25	
Glu (uncharged)	P1		−11	−11
Arg	N0−Qd	N0: Cβ−Cγ−Cδ−Nε	< −25	
Arg (uncharged)	N0−P4	Qd/P4: Cζ−Nω1−Nω2	−23	−25
Lys	C3−Qd	C3: Cβ−Cγ−Cδ	< −25	
Lys (uncharged)	C3−P1	Qd/P1: Cε−Nω	−1	−2
His	SC4−SP1−SP1	SC4: Cβ−Cγ SP1: Cδ−Nε SP1: Nδ−Cε	−19	−20
Phe	SC4−SC4−SC4	SC4: Cβ−Cγ−Cδ1 SC4: Cδ2−Cε2 SC4: Cε1−Cζ	19	17
Tyr	SC4−SC4−SP1	SC4: Cβ−Cγ−Cδ1 SC4: Cδ2−Cε2 SP1: Cϵ1−Cζ−OH	−1	−2
Trp	SC4−SP1−SC4−SC4	SC4: Cβ−Cγ−Cδ2 SP1: Cδ1−Nϵ−Cε1 SC4: Cε2−Cζ2 SC4: Cε1−Cω	12	9

[a] The mapping scheme is reported only for amino acid side chains consisting of more than one CG particle. [b] For the C1 and C2 particle types of the amino acids, the interaction with Q particles has been modified from the standard MARTINI force field. In order to avoid clashes between these particle pairs, the Lennard-Jones parameter σ has been restored from 0.62 nm to the standard value of 0.47 nm.

The most appropriate choice of particle types for the amino acids was assessed from a comparison between simulation results and experimental measurements of the water/oil partitioning coefficients of the amino acid side-chain analogues. The partitioning behavior and amino acid mapping are summarized in Table 1. Simulation data are calculated from equilibrium densities of low concentrations of CG beads dissolved in a water/butane two-phase system. The free energy of partitioning between oil and aqueous phases, $\Delta G^{\mathrm{oil/aq}}$, was obtained from the equilibrium densities ρ of CG particles in both phases:

$$\Delta G^{\mathrm{oil/aq}} = kT \ln\left(\frac{\rho_{\mathrm{oil}}}{\rho_{\mathrm{aq}}}\right) \qquad (7)$$

The equilibrium densities can be obtained directly from a long MD simulation of the two-phase system in which small

Table 2. Backbone Particle Type in Different Kinds of Secondary Structure[a]

backbone	coil bend free	helix	helix (N-terminus/C-terminus)	β-strand turn
backbone	P5	N0	Nd/Na	Nda
Gly	P5	N0	Nd/Na	Nda
Ala	P4	C5	N0	N0
Pro	Na	C5	N0/Na	N0

[a] Both glycine and alanine have no side chain.

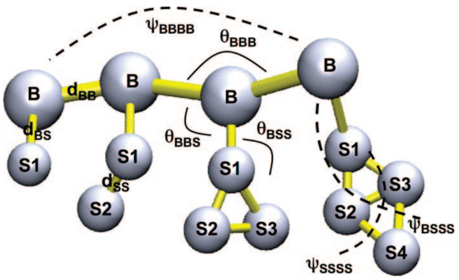


Figure 2. Schematic representation of the four different geometrical classes of amino acids, consisting of either one, two, three, or four beads for the side chain (plus a backbone bead). Intra- and interamino acid bonded potentials are indicated. Backbone beads are indicated by "B" and side-chain beads by "S."

amounts (around 0.01 mol fraction proved sufficient to be in the limit of infinite dilution) of the target substance are dissolved. With the CG model, simulations can easily be extended into the multi-microsecond range, enough to obtain statistically reliable results to within 1 kJ/mol for most particle types. The experimental data[40,41] are for partitioning between water and cyclohexane. Both the simulation and the experimental data are obtained at 300 K. The experimental values could be reproduced to within 2 *kT*, except for the charged amino acids for which no experimental data exist. For some amino acids, especially those consisting of multiple CG beads, more than one assignment scheme would lead to similar partitioning free energies. In those cases, the results obtained for the potential of mean force (PMF) calculations (see section 3.1) were used to select the optimal assignment.

2.3. Parameterization of Bonded Interactions. In Figure 2, the intra amino acid bonded potentials are indicated for the different geometric classes of amino acids (containing either one, two, three, or four side-chain beads plus one backbone bead). The bond lengths, bond angles, dihedral angles, and their respective force constants, collectively referred to as the bonded parameters, were obtained from distributions derived from the PDB. We chose a representative subset of approximately 2000 proteins from the PDB as the basis set for our parametrization. The secondary structure of every residue of these proteins was determined using the program DSSP.[42] Using the center of mass of the atoms representing each coarse-grained bead, we calculated the distributions of the bond lengths, bond angles, and dihedral angles, as shown in Figure 2, for all combinations of amino acids and secondary structures. To ensure that the basis set was truly representative of the entire PDB, we calculated some of the distributions for a different subset of the PDB and found the results to be virtually identical. We also calculated some of the distributions for a membrane protein subset (~200 proteins) and again found the results to be nearly indistinguishable from the original basis set. Additional approaches to calculate the distributions were also attempted, such as using a representative atom instead of the center of mass of the coarse-grained bead. We found that using the center of mass was the most appropriate and robust approach, in line with the general coarse-graining philosophy to represent groups of atoms by an effective particle positioned at their center of mass. After the distributions were obtained, simulations were performed on short test peptides, with different sequences and secondary structure characteristics, and also on larger proteins. All of the bonded parameters were optimized by matching the PDB distributions of the bonds angles and dihedrals with the distributions obtained from the simulations, using an iterative procedure. Representative PDB distributions are shown in Figure 3.

The characteristics of the PDB distributions enabled us to make some approximations, so that the number of tunable parameters was kept at a manageable number without compromising the accuracy. The DSSP definition includes eight secondary structures, namely, helix, extended, bend, turn, beta, 3_{10}-helix, π-helix, and unstructured (random coil). All beta structures were approximated as extended, while the 3_{10}-helix and π-helix were approximated as α-helices. The backbone−backbone bond lengths were all set to be 0.35 nm irrespective of secondary structure. The backbone parameters, that is, the bonds, angles, and dihedrals involving only backbone beads, were set to be dependent on the secondary structure of the beads but independent of the amino acid. Backbone−side-chain (and side-chain−side-chain, where appropriate) bond lengths and force constants were amino acid dependent, but independent of the secondary structure. Backbone−backbone−side-chain and backbone−side-chain−side-chain bond angles and force constants were independent of both the secondary structure and amino acid. Table 3 summarizes the backbone bonded parameters. The force constants in Table 3 correspond to cases where all of the beads involved have the same secondary structure. When a backbone bonded parameter (either a bond or an angle) involves beads with more than one type of secondary structure, the weaker force constant is used. Dihedral angles were imposed only when all four interacting beads had the same secondary structure (either helix or extended). Table 4 summarizes the bond lengths and corresponding force constants for all of the side chains. Table 5 summarizes the bond angles for the side chains.

2.4. Simulation Parameters. The simulations described in this paper were performed with the GROMACS simulation package, version 3.3.1.[43] The topologies, parameters, and example input files of the applications described in this paper are available at http://md.chem.rug.nl/~marrink/coarsegrain.html. Scripts to generate topologies for arbitrary proteins are also downloadable. The general simulation parameters used in the applications described below are as follows. The temperature for each group (lipids, water, and proteins) was kept constant using the Berendsen temperature coupling algorithm[44] with a time constant of 1 ps. Semi-isotropic

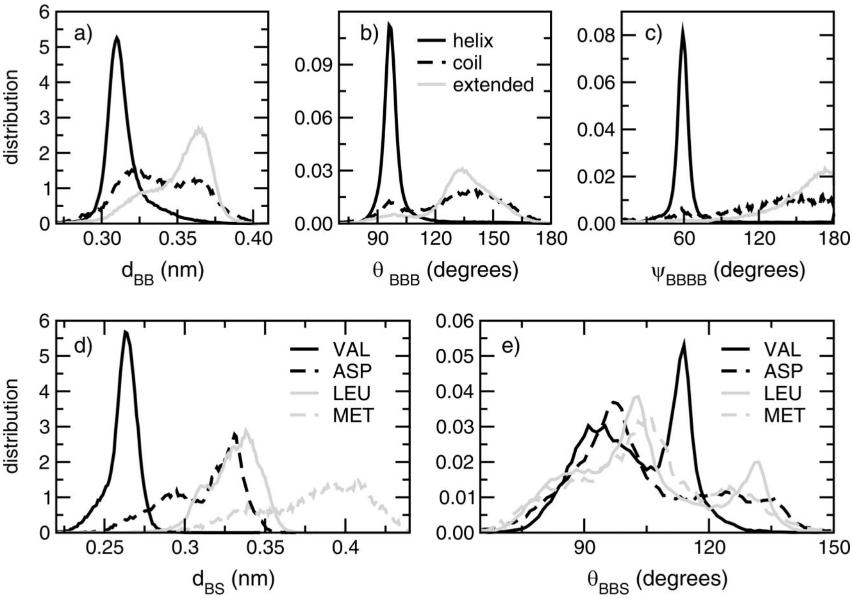


Figure 3. Representative distributions from the PDB. (a) Backbone−backbone bond distributions for helices, random coils, and extended configurations. (b) Backbone−backbone−backbone angle distributions for helices, random coils, and extended configurations. (c) Backbone−backbone−backbone−backbone dihedral angle distributions for helices, random coils, and extended configurations. (d) backbone−side-chain bond distributions for selected aminoacids. (e) Backbone−backbone−side-chain angle distributions for selected amino acids.

Table 3. Backbone Bonded Parameters

backbone	d_{BB} (nm)	K_{BB} (kJ nm^{-2} mol^{-1})	θ_{BBB} (deg)	K_{BBB} (kJ mol^{-1})	ψ_{BBBB} (deg)	K_{BBBB} (kJ mol^{-1})
helix	0.35	1250	96[a]	700	60	400
coil	0.35	200	127	25		
extended	0.35	1250	134	25	180	10
turn	0.35	500	100	25		
bend	0.35	400	130	25		

[a] θ_{BBB} = 98° when Proline is in the helix; K_{BB} = 100 kJ mol^{-1}.

Table 4. Equilibrium Bond Length and Force Constants for Each Amino Acid Side Chain

side chain	d (nm)	K (kJ nm^{-2} mol^{-1})
Leu	0.33	7500
Ile	0.31	constraint
Val	0.265	constraint
Pro	0.30	7500
Met	0.40	2500
Cys	0.31	7500
Ser	0.25	7500
Thr	0.26	constraint
Asn	0.32	5000
Gln	0.4	5000
Asp	0.32	7500
Glu	0.4	5000
Arg d_{BS}	0.33	5000
Arg d_{SS}	0.34	5000
Lys d_{BS}	0.33	5000
Lys d_{SS}	0.28	5000
His d_{BS}	0.32	7500
His d_{SS}	0.27	constraint
Phe d_{BS}	0.31	7500
Phe d_{SS}	0.27	constraint
Tyr d_{BS}	0.32	5000
Tyr d_{SS}	0.27	constraint
Trp d_{BS}	0.3	5000
Trp d_{SS}	0.27	constraint
Cys−Cys d_{S-S}	0.39	5000

pressure coupling was applied using the Berendsen algorithm,[44] with a pressure of 1 bar independently in the plane of the membrane and perpendicular to the membrane. A time constant of 5.0 ps and a compressibility of 4.5×10^{-5} bar^{-1} was used. Bond lengths in aromatic amino acid side chains and the backbone−side-chain bonds for Val, Ile, and Thr were constrained with the LINCS algorithm to avoid numerical instabilities arising from fast fluctuations. Due to the use of shifted potentials, the neighbor list can be updated every 10 steps using a neighbor list cutoff equal to r_{cut} = 1.2 nm. For reasons of computational efficiency, the mass of the CG beads is set to 72 amu (corresponding to four water molecules) for all beads, except for beads in ring structures, for which the mass is set to 45 amu. Using this setup, the systems described in this paper can be simulated with an integration time step of 25 fs, which corresponds to an effective time of 100 fs. In the remainder of the paper, we will use an effective time rather than the actual simulation time unless specifically stated. The CG dynamics are faster than the all-atom dynamics because the CG interactions are much smoother compared to atomistic interactions. The effective friction caused by the fine-grained degrees of freedom is missing. On the basis of comparison of diffusion constants in the CG model and in atomistic models, the effective time sampled using CG interactions is 3–8-fold larger. When interpreting the simulation results with the CG model, a standard conversion factor of 4 is used, which is the effective speedup factor in the diffusion dynamics of CG water compared to real water. The same order of acceleration of the overall dynamics is also observed for a number of other processes, including the permeation rate of water across a membrane,[22] the sampling of the local configurational space of a lipid,[45] and the aggregation rate of lipids into bilayers[22] or vesicles.[43] However, the speedup factor might be different in other systems or for other processes. Particularly for

Table 5. Equilibrium Angles, Improper Dihedral Angles and Force Constants for Side Chains

side chain	θ (deg)	K (kJ mol^{-1})
θ_{BBS} (all)	100	25
θ_{BSS} (Lys, Arg)	180	25
θ_{BSS} (His, Tyr, Phe)	150	50
θ_{BSS} (Trp)	90, 210	50, 50

side chain	ψ (deg)	K (kJ rad^{-2} mol^{-1})
ψ_{BSSS} (His, Tyr, Phe)	0	50
ψ_{BSSS} (Trp)	0, 0	50, 200

protein systems, no extensive testing of the actual speedup due to the CG dynamics has been performed, although protein translational and rotational diffusion was found to be in good agreement with experimental data in simulations of membrane-embedded rhodopsins using a preliminary version of our CG model.[36] In general, however, the time scale of the simulations has to be interpreted with care.

3. Results

3.1. Partitioning of Amino Acid Side Chains in a DOPC Bilayer. The PMF was calculated for each side chain as a function of the distance from a dioleoylphosphatidylcholine (DOPC) bilayer. The results are compared to PMFs obtained at identical conditions and calculated using an analogous procedure from atomistic simulations performed by MacCallum et al.[46,47] The DOPC bilayer contained 72 lipids and 1200 water particles (corresponding to 67 real waters/lipid; excess hydration was chosen to ensure bulk properties of the aqueous phase). While, in principle, the PMF of each side chain analogue can be calculated from unbiased simulations, we used the umbrella sampling method,[48] because unbiased simulations give poor statistics for most side-chain analogues (results not shown) even at simulation times of tens of microseconds. The biasing potential was added to force the CG side chains to sample the region of interest. A series of 46 separate simulations was performed in which the side-chain analogue was restrained to a given distance from the center of the bilayer by a harmonic restraint on the z coordinate only. The spacing between the centers of the biasing potentials was 0.1 nm, and a force constant of 1000 kJ mol^{-1} nm^{-2} was used in all simulations. In each simulation, we placed two side-chain analogues at a distance of 4.5 nm from each other, in order to improve the sampling at practically no additional computational cost. Each of the simulations was 200-ns-long, for a total of 9.2 μs per PMF. After the simulations were completed, unbiased PMFs were extracted using the weighted histogram analysis method.[49]

Figure 4 shows PMF profiles for all 20 side-chain analogues, grouped into hydrophobic, polar, charged, and aromatic pairs of amino acids. In the CG model, the following amino acid side chains are each represented by the same particle type: leucine and isoleucine, valine and proline, cysteine and methionine, and serine and threonine; this choice seems reasonable because atomistic PMFs for these amino acid side chains are similar. The agreement between atomistic and CG PMFs is excellent for hydrophobic amino acids, showing a decrease in the free energy of the system when the side chains enter the bilayer interior. The free energy difference between bulk water and the center of the bilayer is 15 and 17 kJ/mol, respectively, for atomistic and CG leucines, and is 14 and 15 kJ/mol, respectively, for atomistic and CG valines. A free-energy barrier in the proximity of the lipid headgroup region is present in the PMF profiles of most of the hydrophobic residues. This barrier is also well-reproduced in our model for leucine, isoleucine, and valine, with a difference of less than 1 kJ/mol compared to atomistic simulations. For tryptophan and tyrosine, the barrier is not present in the atomistic profiles and is less than 5 kJ/mol in CG profiles. All aromatic residues show a free-energy minimum in the proximity of the interface region of the bilayer. The agreement with atomistic profiles is very good, and the minima have a free energy of −21 and −15 kJ/mol for tryptophan and tyrosine, respectively. PMF profiles for polar amino acids also show a reasonable agreement between atomistic and CG force fields, with higher energies at the center of the bilayer. Only the charged residues show a relatively large difference between the atomistic and CG representation; in all cases, the free-energy penalty for having the residue inside the membrane is underestimated by the CG force field but is still very high

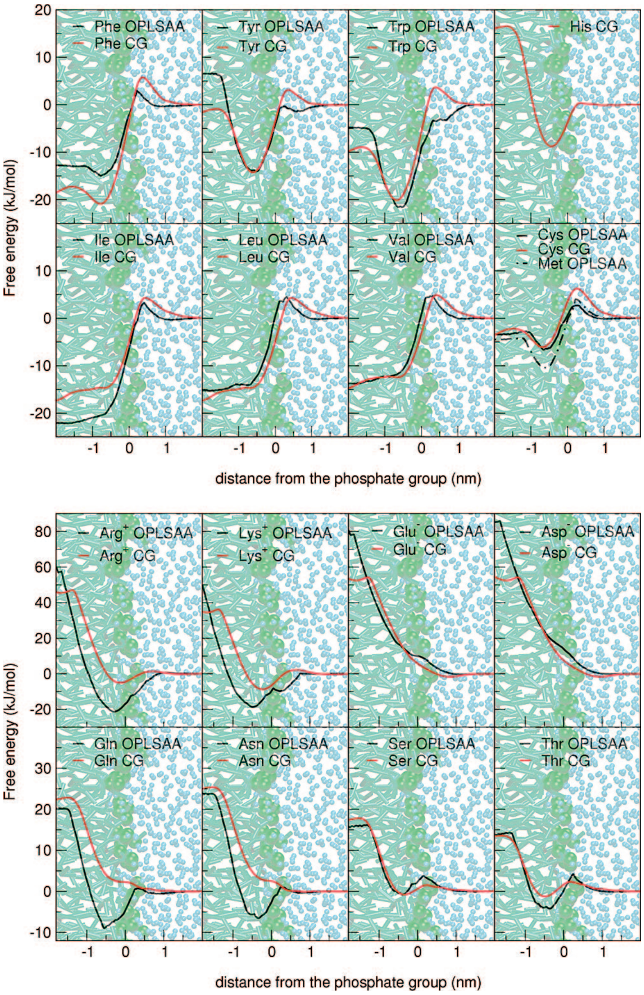


Figure 4. PMF for 16 amino acid side-chain analogues. Results for atomistic simulations (from ref [46]) are shown in black, CG model in red.

Table 6. Association Constant of Leu−Leu and Lys−Glu Residue Pairs Obtained Using the CG and Atomistic Model

	association constant (M^{-1})	
	CG	atomistic
Leu−Leu	3.0 ± 0.3	6.6[a]
Lys−Glu	5.7 ± 0.6	10.8[b]

[a] Yang and Elcock.[54] [b] Thomas and Elcock.[52]

(over 40 kJ/mol for glutamate, aspartate, and arginine, and over 35 kJ/mol for lysine); therefore, the probability of the residues entering the bilayer is negligible. The strong interfacial absorption of the positively charged amino acids predicted from the atomistic simulations is not so well reproduced by the CG force field, and there is room for improvement. Note that solvation free energies in commonly used atomistic force fields show errors up to 8 kJ/mol,[50] with a similar level of accuracy for the free energies of transfer from water to cyclohexane.[51] On the basis of these comparisons, it appears that free energies of partitioning obtained with our coarse-grained model reproduce satisfactorily the results obtained with atomistic models.

3.2. Amino Acid Association Constants. To evaluate the quality of the side-chain−side-chain interactions in the CG force field, association constants between residue pairs were computed and compared to results from atomistic simulations. The pairs Lys/Glu and Leu/Leu, chosen to be representative for salt-bridge and hydrophobic interactions, were placed in a box with 240 CG water particles (960 atomistic water molecules) and run for 4 μs at $T = 300$ K and an isotropic pressure of 1 bar. To avoid freezing of the CG water as a consequence of the small system size (which artificially increases the long-range order), 10% of the water molecules were replaced by antifreeze particles.[23] The association constant K_{ij} between two amino acids i and j can be estimated by[52]

$$K_{ij} = \frac{1}{C} \times \frac{P_{\text{bound}}}{P_{\text{free}}} \tag{8}$$

where $1/C$ is a factor correcting for the concentration of the species in the system, and $P_{(X)}$ is the probability of finding the complex in the X state.[53] Here, $C = 1/(N_A V)$, where N_A is Avogadro's number and V is the volume of the box. The bound and unbound states were differentiated by calculating the solvent accessible surface area (ASA) of the complex. A value of the solvent ASA below a given cutoff indicates that the two residues are in contact, whereas above the same cutoff, the residues are free in solution. The cutoff was chosen from the histogram distribution of the ASA for each simulation at the minimum between the two states. Association constants obtained with the CG model and with atomistic simulations[52,54] are listed in Table 6. The similarity of the values suggests that the contacts observed in the CG model are of reasonable strength. It is especially important that the ratio between hydrophobic and salt-bridge interactions is similar for the all-atom and CG models. Note that the CG model underestimates the strength of these interactions, although only slightly.

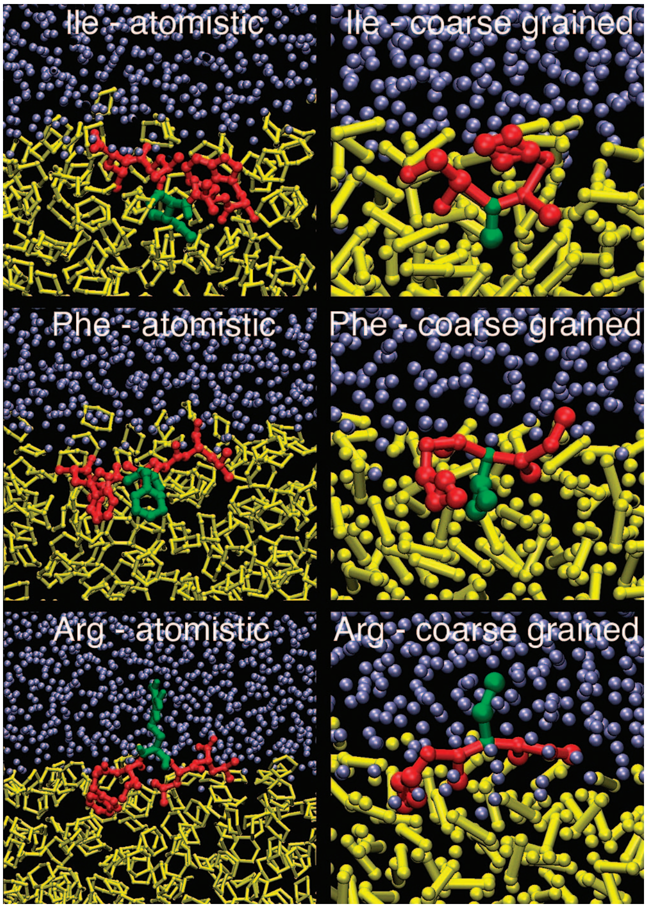


Figure 5. Pentapeptides Ace-WL-X-LL in a water/cyclohexane box in the atomistic representation (left) and WL-X-LL peptide in a water/octane box in the CG representation (right). The peptide is shown in red, with the central residue X in green, water in blue, and the alkane in yellow. Pentapeptides with isoleucine, phenylalanine, and arginine are shown.

3.3. Partitioning and Orientation of Pentapeptides. The series of pentapeptides with sequence Ace-WLXLL was studied by White and Wimley in order to determine an experimental hydrophobicity scale for proteins at membrane interfaces.[55] These peptides are known to partition to the interface region of a zwitterionic membrane, without penetrating into the hydrocarbon core. We simulated a series of 15 pentapeptides with the amino acid sequence Ace-WLXLL, where X is Ala, Arg, Cys, Glu, Gly, His, Ile, Leu, Met, Phe, Pro, Thr, Trp, Tyr, and Val, using both the ffgmx force field as implemented in GROMACS[56] and our CG model. Simulations were carried out in a water/cyclohexane system in the case of the all-atom model, while water/octane was used in the case of the CG model. The Ace group was not present in the CG simulations. No secondary structure was imposed on the peptides. The appropriate particle type P5 was used for the backbone in all peptides (see Table 2). All atomistic simulations were run for 40 ns, while CG simulations were run for 1.5 μs.

We compared the conformation and the partitioning of the whole peptide, as well as the position of residues W1, X3, and L5 relative to the water−alkane interface. Figure 5 shows three of the simulated systems, both at full atomic detail and at the CG level, with isoleucine, phenilalanine, and arginine as central residues. On the basis of backbone angles and head-to-tail distances, all peptides are found mainly in extended conformations, both in atomistic and in CG simulations. In all cases, the peptides partition to the water/alkane interface, consistent with experimental observations. The distributions of the peptides and of single amino acids were evaluated by comparing density profiles. Density profiles for W1, X3, and L5 in CG simulations are similar to those for atomistic simulations for all residues. The peak of the distribution is in the alkane phase for the hydrophobic amino acids and in water for the polar ones, as expected. The average positions relative to the interface show minor differences between atomistic and CG amino acids, lower than 0.1 nm. For the CG model, the density of octane is about 0.77, consistent with the value previously published[22] and close to the experimental value.

3.4. Tilt and Orientation of Helical Peptides. WALP and KALP peptides have sequences consisting of leucine−alanine repeats, flanked by either two tryptophan residues (WALP) or two lysine residues (KALP) at the N terminus and at the C terminus. These model peptides were designed by Killian and co-workers to investigate the effect of hydrophobic mismatch (the difference between the hydrophobic length of the peptide and the hydrophobic width of the lipid) on the properties of zwitterionic membranes.[57,58] The peptides partition into lipid bilayers and assume a transmembrane orientation, with a tilt angle dependent on the extent of mismatch.[59] We investigated the behavior of WALP and KALP peptides, comparing results obtained with the CG model both to experimental data and to simulations performed with various all-atom force fields.

WALP23 in a DOPC Bilayer. In the first set of simulations, we analyzed the tilt and orientation of a WALP23 peptide embedded in a DOPC lipid bilayer containing 72 lipids. We used the same starting conformation (fully α-helical) and orientation (transmembrane, ~0° tilt relative to the bilayer normal) for the atomistic and CG simulations. In atomistic simulations, both the GROMACS (ffgmx)[56] and the OPLS-AA force field[60] were used for the peptide; details can be found elsewhere.[61] Two simulations were carried out with each atomistic force field, and the simulation time was 60 ns for each simulation. One simulation was carried out with the CG force field for 800 ns. For both the atomistic and CG simulations, the temperature was coupled to 300 K and the pressure to 1 bar in the normal and lateral dimensions.

In atomistic simulations, the N-terminal tryptophan side chains are generally found in the proximity of the carbonyl groups of DOPC, while the C-terminal ones are slightly closer to the center of the bilayer. This is consistent with experimental results obtained by fluorescence spectroscopy[62] and mass spectrometry.[63] The tilt angle of the peptide, defined as the angle between the helical axis and the membrane normal, was monitored throughout the simulations. The helix axis was calculated as the first eigenvector of the inertia tensor of the backbone particles. The autocorrelation time of the tilt angle is in the range of tens of nanoseconds, indicating that longer simulations would be required to sample equilibrium distributions. Atomistic simulations of identical systems performed with different initial velocities yielded very different distributions and different averages for the tilt angle, with values of 12 ± 5° and 27 ± 6° in the case of the ffgmx force field and 13 ± 6° and 17 ± 6° for OPLS-AA. The distributions of tilt angles for individual trajectories show the presence of multiple peaks, suggesting that the peptide orientation "jumps" between different arrangements and intermediate orientations are not sufficiently sampled. This indicates that limited sampling impairs predictions of protein orientation in membranes achievable by atomistic simulations. Distributions of the tilt angle for the concatenated trajectories are shown in Figure 6a (atomistic) and b (CG). The average values found in the simulations are significantly larger than those reported from experimental measurements using the GALA method.[59] While the apparent discrepancy can be solved through a different interpretation of the results obtained with the GALA method,[64,65] in the present work, we focus on the comparison between atomistic and CG simulations.

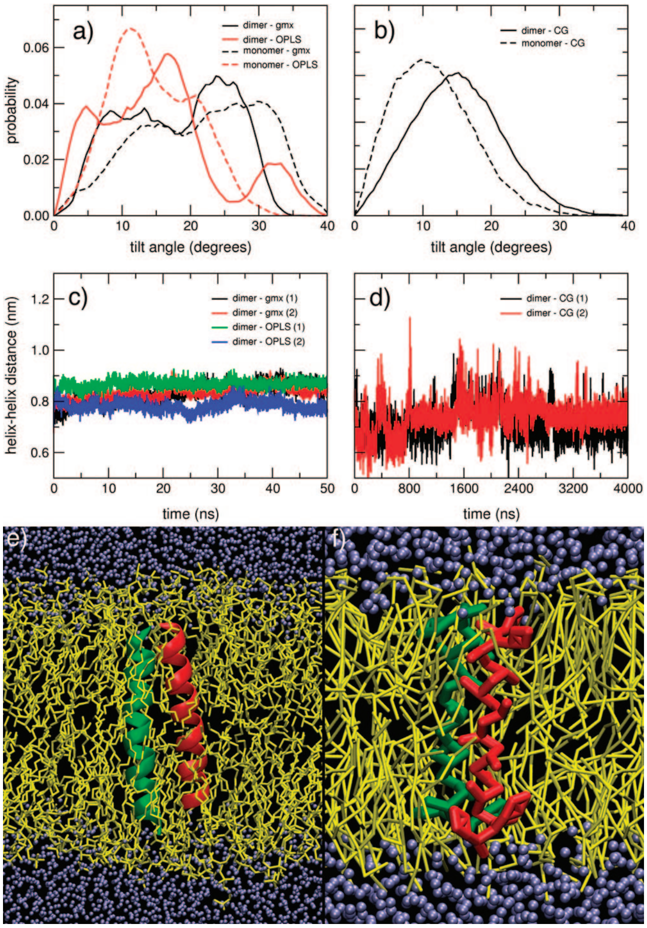


Figure 6. Normalized distributions of tilt angles (angle between the helical axis and the normal to the membrane) in simulations of the WALP23 monomer and dimer in DOPC, using atomistic (a) and CG (b) models. Helix−helix distance in WALP23 dimers in DOPC in atomistic (c) and CG (d) simulations. Atomistic (e) and CG (f) peptides are shown in red and green, with the lipids in yellow and water in blue.

Similarly to the case of atomistic simulations, also in the CG simulations, the average position of the tryptophans is consistent with experiments. Tryptophan side chains are found in the proximity of the GL1 and GL2 particles of DOPC, corresponding approximately to the glycerol and the carbonyl groups of the lipid. Compared to atomistic models, the C-terminal residues appear to be, on average, slightly closer to the hydrophobic portion of the membrane. A broad distribution of tilt angles is observed (see Figure 6b), with an average tilt angle of 11 ± 6°. Contrary to the atomistic simulations, the tilt angles obtained from independent CG simulations were reproduced within one standard error. Given the sampling issues for the atomistic force field, the agreement is reasonable. The autocorrelation time for reorientation of the helical axis is significantly shorter than in the atomistic case (about 3 ns), however. This might point to different kinetic barriers for reorientation of the tryptophan residues near the interface.

KALP Peptides in DLPC and DPPC Bilayers. Recently, a systematic investigation of hydrophobic mismatch, using KALP peptides and PC lipids of different lengths, was performed using atomistic MD simulations.[66] For negative mismatch, when the hydrophobic length of the peptide is smaller than the hydrophobic width of the lipid, a small tilt angle of ~10° was observed. For systematically increasing positive mismatch, a monotonic increase in tilt angles was observed. To compare with these results, CG KALP peptides of different lengths (KALP15, KALP19, KALP23, KALP27, and KALP31) were inserted in DLPC and DPPC membranes consisting of 128 lipids each. The temperature of 310 K and zero surface tension of the membrane match the conditions of the atomistic simulations. Simulations were performed for 200 ns, and the tilt behavior was monitored. The average tilt angles were calculated over the last 100 ns of the simulations. The results are shown in Figure 7, where the tilt angle is shown as a function of mismatch. Snapshots from the atomistic and coarse-grained simulations of KALP31 in DLPC membranes are also shown. Here, the hydrophobic length of the peptide is the distance between the backbone atoms of the first and the last leucine atoms of the peptide, and the hydrophobic width of the lipid is the average distance between the first hydrophobic bead of the lipids in the two leaflets, and the hydrophobic mismatch is the difference between the two. The CG simulations match the trend of the atomistic simulations remarkably well, showing a small tilt angle under negative mismatch and a monotonic increase in tilt angles under positive mismatch.

3.5. Helix−Helix Interactions. Recent experimental evidence suggests that the WALP23 aggregates in DOPC lipid bilayers and forms oligomers of small size if the peptide/lipid ratio is higher than 0.04.[62] We simulated WALP23 dimers in DOPC lipid bilayers in order to compare their stability in atomistic and CG models. All simulation conditions and methodology were identical to the case of the monomers. We used the same two protein force fields, namely, GROMACS (ffgmx) and OPLS-AA, as in simulations reported above. Two simulations were run for each atomistic force field, using the same starting structures and different seed numbers for the initial velocities. We also carried out two simulations using the CG force field. The simulation time was 50 ns for each atomistic simulation and 4 μs for each CG simulation. An antiparallel arrangement of the helices was adopted in all of the simulations. This orientation has been proposed to be favored over the parallel one due to favorable electrostatic interactions between α-helix backbone atoms.[62] The simulated systems are represented in Figure 6 (panels e and f).

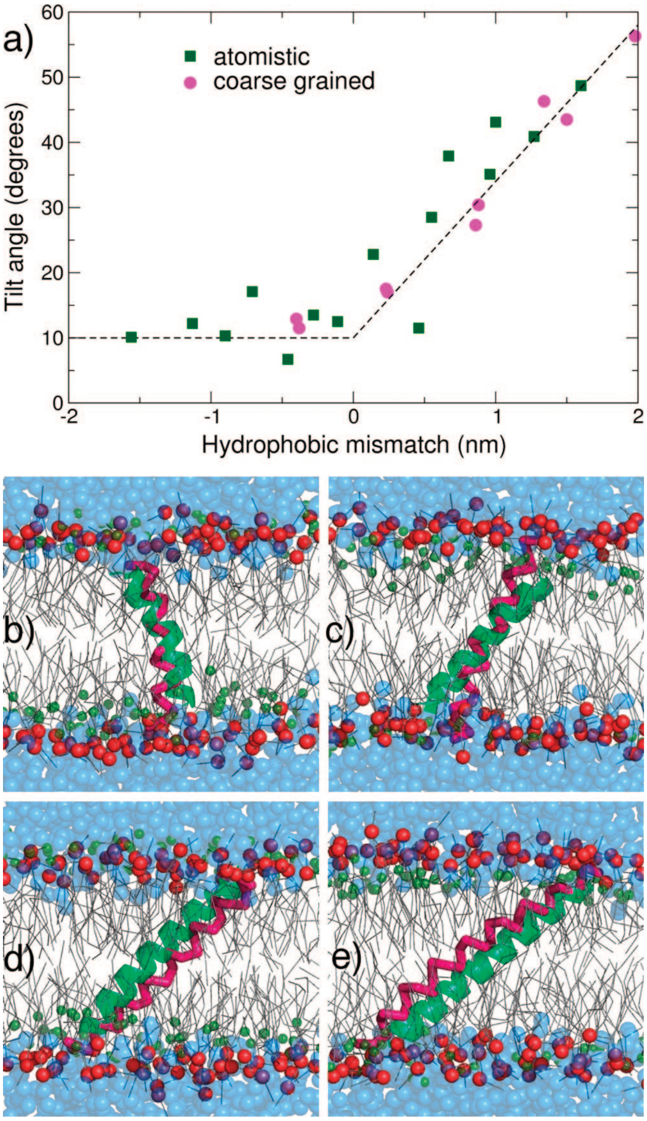


Figure 7. Coarse-grained and atomistic simulations of KALP peptides. (a) The tilt angles are shown as a function of hydrophobic mismatch for the atomistic and CG simulations. In panels b−e, snapshots at the end of the CG simulations are shown. Water is shown as blue spheres, the phosphate group of the lipids as red spheres, the peptide as a pink backbone trace, and the lipids as grey lines. In each of the panels, the peptide from the corresponding atomistic simulation, shown as a green helix, is overlaid on top of the CG peptide for comparison. The phosphorus atoms of the lipids from the atomistic simulations are also show for reference: (b) KALP19 peptide in DLPC lipids, (c) KALP23 peptide in DLPC lipids, (d) KALP27 peptide in DLPC lipids, and (e) KALP31 peptide in DLPC lipids.

In order to assess the stability of the dimers, we monitored the distance between the center of mass of the helices as a function of simulation time (see Table 7 and Figure 6). Helix−helix distances of 0.83 and 0.75 nm were found in the atomistic and CG simulations, respectively, indicating that the dimers are stable in both cases. As in the monomer simulations, also in this case, the peptides are displaced relative to the center of the membrane, with the C-terminal

Table 7. Interhelical Distance and Tilt Angles in the Simulations of the WALP23 Peptide Dimer in DOPC, in Both Atomistic and CG Simulations

simulation	helix−helix distance[a]	tilt angle[b] (helix1)	tilt angle[b] (helix2)
GMX (1)	0.84 ± 0.03°	13.5 ± 4.5°	25.2 ± 4.5°
GMX (2)	0.83 ± 0.04°	9.0 ± 3.5°	24.0 ± 3.5°
OPLS-AA (1)	0.86 ± 0.02°	5.8 ± 3.5°	16.1 ± 4.1°
OPLS-AA (2)	0.78 ± 0.02°	18.9 ± 3.1°	32.3 ± 3.0°
CG (1)	0.71 ± 0.06°	13.6 ± 6.3°	16.0 ± 6.4°
CG (2)	0.70 ± 0.05°	13.7 ± 5.8°	15.4 ± 5.6°

[a] Distance between the center of mass of backbone atoms, in nanometers, ± standard deviation. [b] Angle between the helical axis and the bilayer normal, in degrees, ± standard deviation.

tryptophan side chains closer to the membrane interior. The distributions of tilt angles are shown in Figure 6 (panels a and b). Similarly to the monomer case, also for the dimers, the distributions show multiple peaks for atomistic simulations, highlighting insufficient sampling. On the contrary, distributions obtained with the CG force field are well-converged, with average values for the helix−helix distance and the tilt angles reproduced (within one standard error) in two independent simulations. We notice that the average values of the tilt angle are lower in CG simulations, both for the monomer and for the dimer simulations.

3.6. Partitioning of Hydrophobic Peptides in Lipid Bilayers. Most peptides that partition into lipid membranes exhibit an equilibrium between the transmembrane and the interfacial orientation. Hydrophobic peptides, of lengths comparable to the hydrophobic widths of the lipid bilayer, assemble predominantly into a transmembrane orientation. However, either decreasing the length of the peptide or reducing its hydrophobicity will increase the fraction of the peptides with the interfacial orientation. Hydrophobic polyleucine peptides, with their termini capped by lysines, are highly helical and partition into a predominantly transmembrane orientation in phospholipid bilayers. Substituting some of the leucines with alanines reduces the hydrophobicity of the peptide and should increase the fraction of interfacially bound peptides. In principle, very long simulations should capture the transitions between the transmembrane and interfacial orientations of the peptides. However, even with the CG model, extremely long simulations are required to adequately sample the transitions and obtain meaningful estimates of the transmembrane and interfacially bound fractions. Alternatively, a large number of self-assembly simulations,[67] where a peptide is placed in a random mixture of lipids and water and quenched, can yield a reliable estimate of the transmembrane and interfacially bound fractions, which can be used to calculate a partitioning free energy.

We performed systematic self-assembly simulations of hydrophobic lysine-terminated, polyleucine−alanine peptides in DPPC bilayers. First, six polyleucine peptides, KK(LLLL)$_n$KK were constructed, with n ranging from 1 to 6. Then, the following alanine mutants were created for each value of n: KK(LLLA)$_n$KK, KK(LALA)$_n$KK, KK(LAAA)$_n$-KK, and KK(AAAA)$_n$KK. This gave us a total of 30 peptides with varying lengths and degrees of hydrophobicity. For each of the 30 peptides, 200 independent self-assembly simulations were performed, giving a total of 6000 simulations. In each of the simulations, a single peptide was inserted in a random orientation into a cubic box containing a mixture of 128 DPPC lipids and 1500 CG water molecules in a random arrangement. Then, a self-assembly simulation was performed at 323 K using anisotropic pressure coupling. For this system size, and lipid/water ratio, a bilayer typically forms in about 20 ns. We carried out all of the simulations for 200 ns, providing sufficient time for bilayer self-assembly to occur and for the peptide to partition either to the membrane interior or at the interface or in the water phase. Occasionally, the self-assembly simulation leads to a non-bilayer phase, typically due to one dimension of the simulation cell shrinking rapidly because of the anisotropic pressure coupling. However, this was observed for <2% of all the cases, and these simulations were not included in the partitioning statistics. For all of the simulations which lead to a bilayer phase (>98% of the simulations), the coordinate at 200 ns was visualized and the partitioning state of the peptide was noted. A total of 1.2 ms of data were generated from these simulations.

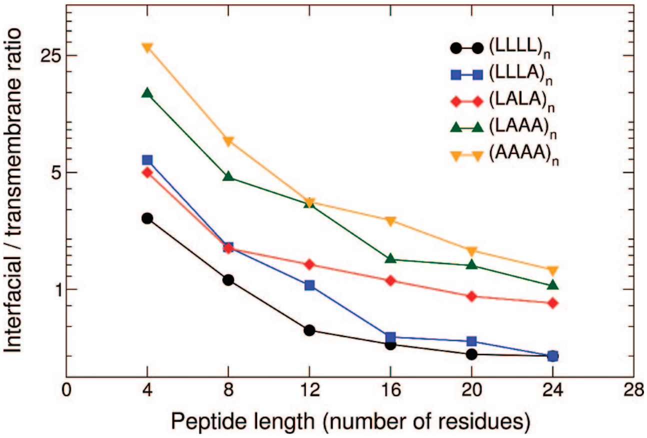


Figure 8. Partitioning behavior of KK-L$_m$A$_n$-KK peptides of varying lengths and hydrophobicity.

We show the interfacial/transmembrane partitioning fraction as a function of peptide length in Figure 8. A priori, we should expect the interfacial fraction to increase as the peptide length is reduced and the alanine content is increased. This is confirmed by our simulations, as seen in Figure 8. It is remarkable that the trends one would expect upon increasing the peptide length and hydrophobicity are faithfully generated. We also observed that the shorter and less-hydrophobic peptides partition into the water phase in significant numbers, as one would expect. The statistics on partitioning into the water phase are too limited to draw more quantitative conclusions. Our results show that the CG model is sensitive enough to capture even minor mutations in the peptide sequence.

3.7. Transmembrane Pores Formed by Antimicrobial Peptides. AMPs are short, cationic, amphipathic peptides that interact with the lipid component of the cell membranes. Magainin-H2 is one of the most well-characterized AMPs, using experimental studies.[68,69] Several biophysical studies have suggested that, at low concentrations, the peptides adopt a surface orientation at the lipid/water interface and, at higher peptide concentrations, the peptides can form toroidal transmembrane pores.[68] However, the exact size and shape

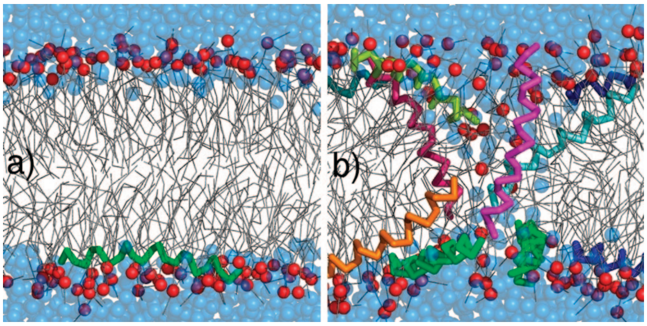


Figure 9. Snapshots of the surface partitioning of an antimicrobial peptide, magainin, at a low concentration (left) and the formation of a toroidal pore at high concentrations (right). In these figures, the water molecules are shown as blue spheres, lipid molecules as grey lines, and the lipid phosphate groups as red spheres. The backbone traces of the peptides are shown as sticks (green on the left and several different colors on the right).

of the toroidal pores is still unclear. We performed self-assembly simulations of single and multiple magainin-H2 peptides using the CG model. The peptides were placed in a random mixture of 128 DPPC lipids and 1500 water molecules. The simulations were performed at 323 K and a pressure of 1 bar. Single peptides almost always partition into a surface-bound orientation. As the number of peptides is increased, with the number of lipids kept constant, some of the peptides partition in a transmembrane orientation. At even higher peptide/lipid (P/L) ratios (5/128 and higher), the self-assembly simulations lead to stable toroidal pores. In Figure 9, we show the final snapshots from two simulations. Figure 9a shows a single peptide bound to the lipid/water interface. Figure 9b shows a toroidal pore, stabilized by the peptides. This structure matches the structures obtained from atomistic simulations of Leontiadou et al.,[70] from which it was concluded that the toroidal pores are rather disordered. In agreement with these atomistic simulations, the structure of the pores formed in the CG simulations differs significantly from current idealized models of a toroidal pore. Instead of multiple peptides lining the pore in a transmembrane orientation, we find typically only one or two peptides near the pore center. The remaining peptides lay close to the edge of the pore, maintaining a predominantly parallel orientation with respect to the membrane. The CG model is thus capable of reproducing the structure of such toroidal pores, arising from the complex interplay between lipids and peptides. Note, in the CG model, the helical conformation of the peptides is enforced by the use of dihedral angle potentials, whereas in the atomistic simulations, significant unfolding occurs. Apparently, unfolding is not a prerequisite for pore formation.

4. Discussion

The potential range of applications of the CG protein model is very broad. Although the test cases shown in this paper involve small peptides only, the model is suited to applications of proteins in general. Processes such as protein aggregation, the action of antimicrobial peptides, protein-induced membrane fusion and fission, ligand binding, and possibly large-scale motions in proteins are amenable to simulation on length and time scales far beyond those feasible with all-atom models. Yet, in contrast to many CG protein models, the resolution at the level of individual amino acids is retained. Using preliminary versions of this force field, the self-assembly of membrane-embedded rhodopsins has already been simulated, for instance,[36] showing that the aggregation is due to a subtle interplay between lipid-mediated long-range attraction and short-range optimization of direct protein–protein contacts. Applications from other groups using similar albeit somewhat simpler models show applications to lipoprotein particles[33] and to a variety of membrane proteins.[34,35]

There are, however, certain important limitations which should be kept in mind. First of all, one has to be aware that secondary structure transformations are not modeled in the current parametrization. The secondary structure is essentially fixed by using angle and dihedral potential energy functions, allowing discrimination between various secondary structure elements but preventing realistic transitions between them. Processes in which the folding and unfolding of secondary structures are playing a substantial role are therefore not suitable for modeling with our current CG force field. Movement of secondary structure elements with respect to each other is possible, however, and is shown to be quite realistic in a recent application of the gating of a membrane-embedded mechanosensitive channel[37] and of voltage-gated potassium channels.[38] Second, the model has been parametrized for the fluid phase. Properties of solids, such as crystal packing, are not expected to be accurate. This might also affect the packing of side chains buried inside proteins, which are somehow in between a fluid and a crystal state. Furthermore, the parametrization is based on free energies. The inherent entropy loss on coarse graining is necessarily compensated for by a reduced enthalpy term. The enthalpy/entropy balance of many processes is therefore biased when modeled at the CG level. Consequently, the temperature dependence is affected, although not necessarily weaker. As is true for any force field, applications outside the temperature range used for parametrization (~270–330 K) have to be considered with care. Another limitation of our CG model, and perhaps of most coarse-graining approaches, is the correct modeling of the partitioning of polar and charged compounds into a low dielectric medium. Because of the implicit screening, the interaction strength of polar substances is underestimated in nonpolarizable solvents. Applications involving the formation of polar/charged complexes in a nonpolar environment are especially prone to being affected. For example, it has been shown in atomistic simulations that charged residues remain hydrated when they are dragged into a lipid bilayer.[46,71,72] In our coarse-grained representation, these residues lose their hydration shell at about 0.7 nm from the center of the bilayer. The difference in hydration leads to a difference between atomistic and CG free-energy profiles. In CG simulations, the free energy of the system increases as charged residues penetrate the lipid bilayer, as long as they are hydrated, and remains flat in the central portion of the membrane. In atomistic profiles, on the other hand, the free energy increases until the residues reach the center of the bilayer. Because of these differences, we suggest that our CG model would probably show artifacts in simulations of the movement of polylysine or arginine domains into lipid bilayers, as is expected for the functioning of certain ion channel proteins. The same is true, in principle, for the action of antimicrobial peptides. Here also, charged residues cross the membrane in one way or another. Despite this potential limitation, in our simulations of magainin-H2, the presence of aqueous pores, in which the charged residues are solvated, makes the problem disappear and explains why we were able to get realistic results. Apart from the implicit screening in the CG model, the neglect of long-range electrostatic forces poses a further potential limitation. Pairwise interactions beyond 1.2 nm (between two and three CG beads away) are not taken into account. In principle, long-range electrostatic interactions could be added to the CG model, in similar ways as is done in atomistic simulations. One has to realize that a modification of the electrostatic interaction scheme will affect other system properties.

Finally, we would like to stress that the current MARTINI model for peptides and proteins is a very general model, designed to be applicable to any class of protein. For any particular application at hand, one could improve the parametrization as required. The bonded interactions are easily fine-tuned on the basis of comparison to either experimental data or to atomistic models. Another promising approach is to use elastic-network models on top of the CG parametrization to mimic the structure and dynamics of a particular native or non-native state.[73] Resolution exchange strategies (i.e., simulations in which CG and all-atom models are combined[74–77]) are promising approaches to further enhance the accuracy and applicability of CG models such as the MARTINI protein model described here.

5. Conclusions

In this paper, we presented an extension of the MARTINI force field to peptides and proteins, enabling simulations of protein systems in the presence of lipids and surfactants at a coarse-grained level. The model allows for a speedup of biomolecular simulations by approximately 3 orders of magnitude compared to traditional all-atom approaches. Importantly, resolution at the level of individual amino acids is retained, and solvent is explicitly taken into account. The protein force field has been parametrized following the same philosophy as the lipid force field. Nonbonded interactions were based on experimental thermodynamic data available for each amino acid. Bonded parameters were derived systematically from distributions of bond lengths, angles, and dihedrals in the Protein Data Bank, allowing for realistic protein conformations to be reproduced. Numerous tests have been performed to validate the choice of parameters. Partitioning of all amino acid side chains in a DOPC lipid bilayer, as well as amino acid association constants, shows good agreement with atomistic simulations. We also studied the partitioning and orientation of numerous model peptides in lipid bilayers: a series of 15 pentapeptides, WALP, KALP, and 30 polyleucine—alanine peptides with different hydrophobicity. Comparison with atomistic simulations and experimental results for all of these model systems demonstrates that our CG force field reproduces the structural and dynamic features of protein–lipid interactions and captures the effect of mutations in the peptide sequence. Finally, the formation of hydrophilic (toroidal) pores in membranes by magainin indicates the great potential of the model for the study of the mechanism of action of antimicrobial and pore-forming peptides, as well as protein aggregation and the effect of peptides and proteins on the properties of biological membranes.

Acknowledgment. This research has been supported by the Natural Science and Engineering Research Council (NSERC). L.M. is an Alberta Heritage Foundation for Medical Research (AHFMR) postdoctoral fellow; D.P.T. is an AHFMR Senior Scholar and CIHR New Investigator. Calculations were performed in part at the WestGrid facilities. S.J.M. acknowledges helpful discussions with Santi Esteban-Martin and Durba Sengupta.

References

(1) Ash, W. L.; Zlomislic, M. R.; Oloo, E. O.; Tieleman, D. P. Computer simulations of membrane proteins. *Biochim. Biophys. Acta* **2004**, *1666* (1–2), 158–189.

(2) Go, N.; Taketomi, H. Respective roles of short-range and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75* (2), 559–563.

(3) Miyazawa, S.; Jernigan, R. L. Estimation of effective inter-residue contact energies from protein crystal-structures - quasi-chemical approximation. *Macromolecules* **1985**, *18* (3), 534–552.

(4) Das, P.; Matysiak, S.; Clementi, C. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (29), 10141–10146.

(5) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des.* **1997**, *2* (3), 173–181.

(6) Tama, F.; Wriggers, W.; Brooks, C. L. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Biol.* **2002**, *321* (2), 297–305.

(7) Chan, H. S.; Dill, K. A. Compact polymers. *Macromolecules* **1989**, *22* (12), 4559–4573.

(8) Chan, H. S.; Dill, K. A. Intrachain loops in polymers - effects of excluded volume. *J. Chem. Phys.* **1989**, *90* (1), 492–509.

(9) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253* (5494), 694–8.

(10) 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* (7), 2362–7.

(11) Tozzini, V.; McCammon, J. A. A coarse grained model for the dynamics of flap opening in HIV-1 protease. *Chem. Phys. Lett.* **2005**, *413* (1–3), 123–128.

(12) Arkhipov, A.; Freddolino, P. L.; Imada, K.; Namba, K.; Schulten, K. Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum. *Biophys. J.* **2006**, *91* (12), 4589–4597.

(13) Venturoli, M.; Sperotto, M. M.; Kranenburg, M.; Smit, B. Mesoscopic models of biological membranes. *Phys. Rep.* **2006**, *437* (1–2), 1–54.

(14) Sperotto, M. M.; May, S.; Baumgaertner, A. Modelling of proteins in membranes. *Chem. Phys. Lipids* **2006**, *141* (1–2), 2–29.

(15) Shillcock, J. C.; Lipowsky, R. The computational route from bilayer membranes to vesicle fusion. *J. Phys.: Condens. Matter* **2006**, *18*, S1191–S1219.

(16) Müller, M.; Katsov, K.; Schick, M. Biological and synthetic membranes: What can be learned from a coarse-grained description. *Phys. Rep.* **2006**, *434*, 113–176.

(17) 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* (3), 1778–98.

(18) Smeijers, A. F.; Pieterse, K.; Markvoort, A. J.; Hilbers, P. A. J. Coarse-grained transmembrane proteins: Hydrophobic matching, aggregation, and their effect on fusion. *J. Phys. Chem. B* **2006**, *110* (27), 13614–13623.

(19) 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* (31), 15045–15048.

(20) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett.* **2005**, *95* (21).

(21) Heath, A. P.; Kavraki, L. E.; Clementi, C. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins* **2007**, *68* (3), 646–661.

(22) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760.

(23) Marrink, S. J.; Risselada, H. 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.

(24) Marrink, S. J.; Mark, A. E. Molecular dynamics simulation of the formation, structure, and dynamics of small phospholipid vesicles. *J. Am. Chem. Soc.* **2003**, *125* (49), 15233–15242.

(25) Marrink, S. J.; Mark, A. E. Molecular view of hexagonal phase formation in phospholipid membranes. *Biophys. J.* **2004**, *87* (6), 3894–3900.

(26) Faller, R.; Marrink, S. J. Simulation of domain formation in DLPC-DSPC mixed bilayers. *Langmuir* **2004**, *20* (18), 7686–7693.

(27) Ashbaugh, H. S.; Patel, H. A.; Kumar, S. K.; Garde, S. Mesoscale model of polymer melt structure: Self-consistent mapping of molecular correlations to coarse-grained potentials. *J. Chem. Phys.* **2005**, *122* (10), 104908.

(28) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. A coarse grain model for phospholipid simulations. *J. Phys. Chem. B* **2001**, *105* (19), 4464–4470.

(29) Lyubartsev, A. P. Multiscale modeling of lipids and lipid bilayers. *Eur. Biophys. J. Biophys. Lett.* **2005**, *35* (1), 53–61.

(30) Elezgaray, J.; Laguerre, M. A systematic method to derive force fields for coarse-grained simulations of phospholipids. *Comput. Phys. Commun.* **2006**, *175* (4), 264–268.

(31) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109* (7), 2469–2473.

(32) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676.

(33) 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* (8), 3674–3684.

(34) Bond, P. J.; Sansom, M. S. P. Insertion and assembly of membrane proteins via simulation. *J. Am. Chem. Soc.* **2006**, *128* (8), 2697–2704.

(35) Bond, P. J.; Sansom, M. S. Bilayer deformation by the Kv channel voltage sensor domain revealed by self-assembly simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (8), 2631–6.

(36) (a) 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* (33), 10126–10132. (b) Botelho, A. V.; Huber, T.; Sakmar, T. P.; Brown, M. F. Curvature and hydrophobic forces drive oligomerization and modulate activity of rhodopsin in membranes. *Biophys. J.* **2006**, *91* (12), 4464–4477.

(37) Yefimov, S.; Giessen, E. v. d.; Onck, P.; Marrink, S. Mechanosensitive membrane channels in action *Biophys. J.* **2008**, *94* (8), 2994–3002.

(38) 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* (11), 3277–3282.

(39) Catte, A.; Patterson, J. C.; Bashtovyy, D.; Jones, M. K.; Gu, F.; Li, L.; Rampioni, A.; Sengupta, D.; Vuorela, T.; Niemelä, T.; Karttunen, M.; Marrink, S. J.; Vattulainen, I.; Segrest, J. P. Structure of spheroidal HDL particles revealed by combined atomistic and coarse grained simulations. *Biophys. J.* **2008**, *94* (6), 2306–2319.

(40) 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* (5), 1664–1670.

(41) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Affinities of amino-acid side-chains for solvent water. *Biochemistry* **1981**, *20* (4), 849–855.

(42) Kabsch, W.; Sander, C. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577–2637.

(43) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. Gromacs: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(44) 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.

(45) Baron, R.; Trzesniak, D.; de Vries, A. H.; Elsener, A.; Marrink, S. J.; van Gunsteren, W. F. Comparison of thermodynamic properties of coarse-grained and atomic-level simulation models. *ChemPhysChem* **2007**, *8* (3), 452–461.

(46) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.* **2008**, . in press.

(47) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Partitioning of amino acid side chains into lipid bilayers: results from computer simulation and comparison to experiment. *J. Genet. Physiol.* **2007**, *129* (5), 371–377.

(48) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distribution in Monte Carlo free energy estimation: umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(49) 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* (8), 1011–1021.

(50) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* **2003**, *119* (11), 5740–5761.

(51) Maccallum, J. L.; Tieleman, D. P. Calculation of the water-cyclohexane transfer free energies of neutral amino acid side-chain analogs using the OPLS all-atom force field. *J. Comput. Chem.* **2003**, *24* (15), 1930–1935.

(52) Thomas, A. S.; Elcock, A. H. Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures. *J. Am. Chem. Soc.* **2004**, *126* (7), 2208–2214.

(53) Hunenberger, P. H.; Granwehr, J. K.; Aebischer, J. N.; Ghoneim, N.; Haselbach, E.; vanGunsteren, W. F. Experimental and theoretical approach to hydrogen-bonded diastereomeric interactions in a model complex. *J. Am. Chem. Soc.* **1997**, *119* (32), 7533–7544.

(54) Yang, H. B.; Elcock, A. H. Association lifetimes of hydrophobic amino acid pairs measured directly from molecular dynamics simulations. *J. Am. Chem. Soc.* **2003**, *125* (46), 13968–13969.

(55) Wimley, W. C.; White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **1996**, *3* (10), 842–8.

(56) van der Spoel, D.; Lindahl, E.; Hess, B.; 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*; 1991–2005.

(57) de Planque, M. R. R.; Greathouse, D. V.; Koeppe, R. E.; Schafer, H.; Marsh, D.; Killian, J. A. Influence of lipid/peptide hydrophobic mismatch on the thickness of diacylphosphatidylcholine bilayers. A H-2 NMR and ESR study using designed transmembrane α-helical peptides and gramicidin A. *Biochemistry* **1998**, *37* (26), 9333–9345.

(58) van der Wel, P. C. A.; de Planque, M. R. R.; Greathouse, D. V.; Koeppe, R. E.; Killian, J. A. Effects of hydrophobic mismatch on the interaction between transmembrane α-helical peptides with TRP anchors, and lipid bilayers of phosphatidylcholine and phosphatidnethanolamine. *Biophys. J.* **1998**, *74* (2), A304–A304.

(59) Strandberg, E.; Ozdirekcan, S.; Rijkers, D. T. S.; van der Wel, P. C. A.; Koeppe, R. E.; Liskamp, R. M. J.; Killian, J. A. Tilt angles of transmembrane model peptides in oriented and non-oriented lipid bilayers as determined by H-2 solid-state NMR. *Biophys. J.* **2004**, *86* (6), 3709–3721.

(60) Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential functions for proteins - energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666.

(61) Tieleman, D. P.; MacCallum, J. L.; Ash, W. L.; Kandt, C.; Xu, Z.; Monticelli, L. Membrane protein simulations with a united-atom lipid and all-atom protein model: lipid-protein interactions, side chain transfer free energies and model proteins. *J. Phys.: Condens. Matter* **2006**, *18* (28), S1221.

(62) Sparr, E.; Ash, W. L.; Nazarov, P. V.; Rijkers, D. T.; Hemminga, M. A.; Tieleman, D. P.; Killian, J. A. Self-association of transmembrane alpha-helices in model membranes: importance of helix orientation and role of hydrophobic mismatch. *J. Biol. Chem.* **2005**, *280* (47), 39324–31.

(63) Demmers, J. A. A.; Haverkamp, J.; Heck, A. J. R.; Koeppe, R. E.; Killian, J. A. Electrospray ionization mass spectrometry as a tool to analyze hydrogen/deuterium exchange kinetics of transmembrane peptides in lipid bilayers. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (7), 3189–3194.

(64) Özdirekcan, S.; Etchebest, C.; Killian, J. A.; Fuchs, P. F. J. On the orientation of a designed transmembrane peptide: towards the right tilt angle. *J. Am. Chem. Soc.* **2007**, *129*, 15174–15181.

(65) Esteban-Martin, S.; Salgado, J. The dynamic orientation of membrane-bound peptides: Bridging simulations and experiments. *Biophys. J.* **2007**, [Online] biophysj.107.113043.

(66) Kandasamy, S. K.; Larson, R. G. Molecular dynamics simulations of model trans-membrane peptides in lipid bilayers: a systematic investigation of hydrophobic mismatch. *Biophys. J.* **2006**, *90* (7), 2326–2343.

(67) Esteban-Martin, S.; Salgado, J. Self-assembling of peptide/membrane complexes by atomistic molecular dynamics simulations. *Biophys. J.* **2007**, *92* (3), 903–912.

(68) Lee, M. T.; Hung, W. C.; Chen, F. Y.; Huang, H. W. Many-body effect of antimicrobial peptides: On the correlation between lipid's spontaneous curvature and pore formation. *Biophys. J.* **2005**, *89* (6), 4006–4016.

(69) Ludtke, S. J.; He, K.; Heller, W. T.; Harroun, T. A.; Yang, L.; Huang, H. W. Membrane pores induced by magainin. *Biochemistry* **1996**, *35* (43), 13723–13728.

(70) Leontiadou, H.; Mark, A. E.; Marrink, S. J. Antimicrobial peptides in action. *J. Am. Chem. Soc.* **2006**, *128* (37), 12156–12161.

(71) Monticelli, L.; Robertson, K. M.; MacCallum, J. L.; Tieleman, D. P. Computer simulation of the KvAP voltage-gated potassium channel: steered molecular dynamics of the voltage sensor. *FEBS Lett.* **2004**, *564* (3), 325–332.

(72) Freites, J. A.; Tobias, D. J.; von Heijne, G.; White, S. H. Interface connections of a transmembrane voltage sensor. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (42), 15059–15064.

(73) 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* (3), 593–605.

(74) 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* (2), 192–8.

(75) Christen, M.; van Gunsteren, W. F. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *J. Chem. Phys.* **2006**, *124*, 154106.

(76) Praprotnik, M.; Delle Site, L.; Kremer, K. A macromolecule in a solvent: Adaptive resolution molecular dynamics simulation. *J. Chem. Phys.* **2007**, *126*, 134902.

(77) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. Resolution exchange simulation. *Phys. Rev. Lett.* **2006**, *962*.

CT700324X

JCTC Journal of Chemical Theory and Computation

Multiscale Monte Carlo Sampling of Protein Sidechains: Application to Binding Pocket Flexibility

Jerome Nilmeier*,† and Matt Jacobson‡

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

Received December 10, 2007

Abstract: We present a Monte Carlo sidechain sampling procedure and apply it to assessing the flexibility of protein binding pockets. We implemented a multiple "time step" Monte Carlo algorithm to optimize sidechain sampling with a surface generalized Born implicit solvent model. In this approach, certain forces (those due to long-range electrostatics and the implicit solvent model) are updated infrequently, in "outer steps", while short-range forces (covalent, local nonbonded interactions) are updated at every "inner step". Two multistep protocols were studied. The first protocol rigorously obeys detailed balance, and the second protocol introduces an approximation to the solvation term that increases the acceptance ratio. The first protocol gives a 10-fold improvement over a protocol that does not use multiple time steps, while the second protocol generates comparable ensembles and gives a 15-fold improvement. A range of 50−200 inner steps per outer step was found to give optimal performance for both protocols. The resultant method is a practical means to assess sidechain flexibility in ligand binding pockets, as we illustrate with proof-of-principle calculations on six proteins: DB3 antibody, thermolysin, estrogen receptor, PPAR-γ, PI3 kinase, and CDK2. The resulting sidechain ensembles of the apo binding sites correlate well with known induced fit conformational changes and provide insights into binding pocket flexibility.

Introduction

Sidechain sampling and optimization algorithms, mostly based on a rotamer approximation,[1–5] have been used extensively in modeling proteins, including homology modeling,[6,7] and predicting conformational changes due to ligand binding.[8–10] We have been interested in developing sampling methods for protein sidechains (and, in other work, loops) that generate thermodynamic ensembles of conformations, in contrast to locating the global energy minimum.[11,12] Minimization methods implicitly neglect the effect of entropy on sidechain conformations, and generally cannot distinguish whether sidechains will adopt a single well-defined conformation, or a distribution of conformations. For the many sidechains that are tightly packed in the core of a protein, minimization is an effective approach. For less tightly packed sidechains that display some degree of flexibility, a thermodynamic ensemble becomes a more appropriate description.

Sidechain conformational heterogeneity is important to protein−ligand binding. The ability to accurately predict the flexibility/rigidity of binding site residues would be useful in structure-based drug design.[10,13] For example, a recent paper by Sherman et al.[8] describes a computational method to predict "induced fit" effects upon ligand binding which relies on some advanced knowledge of which sidechains may adopt different conformations upon ligand binding, e.g., from multiple cocrystal structures. We demonstrate here that thermodynamic ensembles of sidechain conformations in apo proteins correlate well with known induced fit conformational changes in various well studied drug targets.

In principle, molecular dynamics sampling methods[10,14] can be used to obtain thermodynamic ensembles for protein binding sites. The main disadvantage is that the timescales required to observe large changes in sidechain conformations can be long relative to the ∼1 fs timesteps employed in

* Corresponding author. E-mail: jerome.nilmeier@ucsf.edu.

† Graduate Group in Biophysics.

‡ Department of Pharmaceutical Chemistry.

10.1021/ct700334a CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/16/2008

atomically detailed molecular dynamics simulations; transitions between sidechain rotamers can take up to μs, which is a known difficulty in binding affinity calculations.[14–16] Monte Carlo sampling[17] can lead to more efficient generation of the complete thermodynamic ensemble, if the trial moves are constructed carefully.

For macromolecules, which contain complex, heterogeneous, and densely packed atomic configurations, construction of efficient trial moves can be a substantial challenge. A variety of both rigorous and nearly rigorous methods have been used[12,18–23] to address this challenge. One common idea among these involves decomposing the degrees of freedom into subspaces that are more manageable, both computationally and conceptually. The most natural decomposition for proteins is between backbone and sidechain degrees of freedom. Future work will incorporate backbone motions, but the current emphasis is on the sidechain sampling.

Another common decomposition is between solvent (water) and solute (protein) degrees of freedom. Here we use an implicit solvent model, which makes it possible to efficiently sample large sidechain conformational changes. By contrast, in explicit solvent, large changes (e.g., across rotamers) are difficult to sample with good acceptance rates because of steric clashes between waters and the sidechain, and the need for the solvent to relax around any new trial conformation. The same steric issues have motivated the use of implicit solvent in molecular dynamics studies as well.[24–26] For this work, the electrostatic solvation term is evaluated with the SGB model[27,28] and the nonpolar solvation energy with the nonpolar (NP) model.[29] The solvation model here was developed for use with the all atom OPLS-AA 2001 forcefield[30] and is implemented in the Protein Local Optimization Program.[31,32] While this model is chosen as a compromise between efficiency and accuracy, it remains the most computationally expensive portion of the energy evaluations. The current effort is to develop a general sampling scheme which allows optimal use of an implicit solvation model in the context of a Monte Carlo scheme. The present application is to sidechain sampling, but can be extended to backbone sampling strategies in a straightforward manner.

The major innovation here in terms of computational methods is the implementation of a multiscale strategy, analogous to methods such as RESPA,[33,34] used in molecular dynamics, to accelerate convergence toward the thermodynamic ensemble. In this approach, certain forces (primarily those due to long-range electrostatics and the implicit solvent model) are updated infrequently, in “outer steps”, while short-range forces (covalent, local nonbonded interactions) are updated at every “inner step”. The theory underlying this approach has been presented previously,[35] and is only briefly reviewed here. The application of a multiscale Monte Carlo approach to sampling proteins in implicit solvent has been presented by Michel et al.,[36] with different implementation details and approximations introduced. Other algorithmic details crucial for speed, including the rapid elimination of conformations with steric clashes, are also described. The resultant method is a practical means to assess sidechain flexibility in ligand binding pockets, as we illustrate with proof-of-principle calculations on six proteins.

Theory and Methods

Configuration Integral. The implicitly solvated[37] macromolecular ensembles of interest can be represented by the following configuration integral:

$$Q = \int \mathrm{d}\mathbf{R}\, \exp(-\beta[A(\mathbf{R})]) \tag{1}$$

where **R** is the set of all Cartesian coordinates of the macromolecule of interest, and

$$A(\mathbf{R}) = U(\mathbf{R}) + G(\mathbf{R}) \tag{2}$$

where $A(\mathbf{R})$ is the sum of the forcefield energy, $U(\mathbf{R})$, and the implicit solvation energy, $G(\mathbf{R})$. The solvation energy is dependent on the Born radii, which are a function of the coordinate state of the macromolecule. In the SGB implementation we use, the Born radii $\boldsymbol{\alpha}(\mathbf{R})$ are computed using surface integrals, and thus are dependent on the global coordinate state **R** of the protein. This calculation can take much longer (roughly 100 times longer in cases studied) than the pairwise energy terms. Some improvements have been gained by updating only local regions of the surface area as needed, and efforts are ongoing in this area to improve the efficiency and accuracy of this model.[38,39]

In general, however, any attempt to optimize sampling would benefit most from evaluating the solvation energy less frequently. While this approach is motivated by computational efficiency, a physical argument can also be made. The Born radii generally vary slowly for relatively small, local conformational changes. The sampling strategies presented are intended to make the best use of these ideas while still generating meaningful ensembles.

Constraints on various degrees of freedom can be introduced to generate a configuration integral q_0 over a smaller subspace by identifying fixed (F) and sampled (S) degrees of freedom, such that $\mathrm{d}\mathbf{R} = \mathrm{d}\mathbf{R}^{<F>}\ \mathrm{d}\mathbf{R}^{<S>}$, and imposing a rigid constraint on the fixed degrees of freedom, yielding

$$q_0 = \int \mathrm{d}\mathbf{R}^{<S>} \exp\left(-\beta A\left[\left(\mathbf{R}^{<S>}|\mathbf{R}_0^{<F>}\right)\right]\right) \tag{3}$$

Following the formulation of Deem,[20] the transformation from Cartesian to torsional coordinates can be made with a Jacobian of unity, if bond lengths and angles are preserved. For the current work, the backbone torsions will be constrained to an initial value of $\boldsymbol{\varphi}_0$, and the fixed sidechains, to an initial value of $\boldsymbol{\chi}_0^{<F>}$. The resulting integral can be recast as

$$q_0 = \int \mathrm{d}\boldsymbol{\chi}^{<S>} \exp\left(-\beta\left[A\left(\boldsymbol{\chi}^{<S>}|\boldsymbol{\varphi}_0, \boldsymbol{\chi}_0\right)\right]\right) \tag{4}$$

where $\boldsymbol{\chi}^{<S>}$ is the set of sidechain torsional coordinates that are sampled. The integral of interest over the subspace can be recast by letting $\mathrm{d}\mathbf{r} = \mathrm{d}\mathbf{R}^{<S>}$ and $A(\mathbf{r}) = A(\boldsymbol{\chi}^{<F>}) = A(\mathbf{R}^{<S>}|\mathbf{R}_0^{<F>})$, yielding the more compact expression:

$$q = \int \mathrm{d}\mathbf{r}\, \exp(-\beta[A(\mathbf{r})] \tag{5}$$

Generation of Trial Configurations. To generate a reversible trial move, a single sidechain i is chosen at random from the list of sampled sidechains and the updated set of torsions is assigned according to

$$\chi_i' = \chi_i + \xi \qquad (6)$$

where χ_i' and χ_i are the trial and previous set of dihedral coordinates, respectively, for sidechain *i*, and ξ is a vector of uniform random variates of the same dimension, for which each value is drawn from the domain $[-d/2, d/2]$. To account for local fluctuations as well as larger fluctuations, the domain size *d* is assigned a value of either 360° or 18° with equal probability. The idea behind the heterogeneous move set is to alternate between large dihedral trial moves that cross local χ wells, and small trial moves, which sample the local χ basin. For the present work, selections from a rotamer library are not incorporated as a trial move, as slight nonuniformities in the distribution of the χ angles of the rotamer library have a quantitative effect on the distributions. As a practical matter, however, a mixture of rotamer and random moves could conceivably be implemented if quantitative energy distributions are not required.

For residues with rotatable polar hydrogen groups (Cys, Ser, Thr, Tyr), the torsional angle that places the hydrogen is also selected randomly when the rotamer state is assigned. Also, the torsions of the amine hydrogens of lysines are sampled. Torsions for methyl hydrogens are not currently sampled.

A hard sphere approximation is invoked, which vastly improves sampling efficiency, while preserving much of the essential physics of the system. This has been shown in liquid systems[40,41] as well as proteins. For the current work, pairs of atoms that are closer than 0.7 times the sum of the Lennard-Jones radii are considered to be sterically disallowed. That is, no energy is computed for sterically disallowed states, because the steric clash will result in high energies and small acceptance probabilities. Cell lists (linked lists) further accelerate the identification of steric clashes, by only checking for clashes between atoms known to be proximal. A series of dihedral perturbations is generated as described until a configuration that is sterically allowed is generated. The resulting configuration is treated as a trial move. For the systems studied, the average number of sterically disallowed moves ranges from 0.5 to 0.75 (see Table 2), which is roughly a 2−4-fold improvement in sampling efficiency, because the CPU time per steric clash evaluation is negligible relative to the energy evaluation.

Multiple Time Step Monte Carlo (MTS-MC). A sampling procedure known as multiple time step Monte Carlo,[35] which was originally developed for Ewald sum calculations,[42] can be used to optimally sample against a potential that can be decomposed into additive components. These components are typically, but not necessarily, short- and long-range contributions to the energy. The algorithm relies on the assumption that the short-range term varies rapidly with respect to the move set, while the long-range term varies more slowly. A related formalism is presented using approximate potentials.[43] Many algorithms use similar ideas, including both molecular dynamics integrators[33,34] and minimization algorithms.[44] Some applications using algorithms that are similar in spirit involve evaluating Ewald sums less frequently in fluid simulations with periodic boundary conditions, sampling of polar fluids,[45] and polarizable water sampling.[46]

While the formalisms in these approaches vary, they can all be thought of as relying on some decomposition of the overall potential to be sampled. The natural choice of decomposition, in general, is into short- and long-range terms, which we denote by subscripts S and L, respectively

$$A(\mathbf{r}) = A_S(\mathbf{r}) + A_L(\mathbf{r}) \qquad (7)$$

The details of the nature of the decomposition of interactions into long and short-range can vary from system to system. A more detailed description of the decomposition for the present case, with proof of detailed balance, is given in the Appendix.

Using the above decomposition, detailed balance can be maintained using the following sampling protocol:

(1) Starting with the configuration $\mathbf{r}_i$, generate a number N_I of inner loop steps, where each step consists of a trial configuration $\mathbf{r}_{k'}$ that is generated reversibly (such as the trial configurations described by eq 6) and accepted according to the following short-range acceptance criterion:

$$\frac{\mathrm{acc}_S(\mathbf{r}_{k'}|\mathbf{r}_k)}{\mathrm{acc}_S(\mathbf{r}_k|\mathbf{r}_{k'})} = \exp(-\beta[A_S(\mathbf{r}_{k'}) - A_S(\mathbf{r}_k)]) \qquad (8)$$

(2) Take the final configuration from the inner loop to be the trial configuration $\mathbf{r}_j$ for the outer loop and apply the long-range acceptance criterion:

$$\frac{\mathrm{acc}_L(\mathbf{r}_j|\mathbf{r}_i)}{\mathrm{acc}_L(\mathbf{r}_i|\mathbf{r}_j)} = \exp(-\beta[A_L(\mathbf{r}_j) - A_L(\mathbf{r}_i)]) \qquad (9)$$

It is important to note that any statistical quantities of interest can only be computed using the outer loop configurations. In all cases where the ratio of acceptance probabilities are given, the Metropolis acceptance criterion is used in practice.

Recasting MTS-MC to Account for Infrequent Born Radii Updates. For the present case, the most costly term to evaluate in the energy is the solvation term, which is due largely to the time intensive step of computing the Born radii, $\boldsymbol{\alpha}(\mathbf{R})$, and we develop a strategy such that the Born radii are not updated in the inner steps. To motivate this method, it is helpful to express the potential in the following form:

$$A(\alpha(\mathbf{R}_m), \mathbf{r}_n) = U(\mathbf{r}_n) + G(\alpha(\mathbf{R}_m), \mathbf{r}_n) \qquad (10)$$

where $\mathbf{r}_n$ is *n*th configuration of the subset of sampled coordinates, $\boldsymbol{\alpha}(\mathbf{R}_m)$ is the set of Born radii which are evaluated based on the coordinates of the *m*th coordinate state $\mathbf{R}_m$ of the entire protein, $U(\mathbf{r}_n)$, and $G(\boldsymbol{\alpha}(\mathbf{R}_m), \mathbf{r}_n)$ is the solvation energy evaluated at the given states. We can further express the energy deviation from the "true" potential, where the Born radii are synchronous with the current coordinate state, in terms of an error potential $\varepsilon(\boldsymbol{\alpha}(\mathbf{R}_m), \mathbf{r}_n)$:

$$\varepsilon(\alpha(\mathbf{R}_m), \mathbf{r}_n) = A(\alpha(\mathbf{R}_n), \mathbf{r}_n) - A(\alpha(\mathbf{R}_m), \mathbf{r}_n) = G(\alpha(\mathbf{R}_n), \mathbf{r}_n) - G(\alpha(\mathbf{R}_m), \mathbf{r}_n) \qquad (11)$$

Thus, the inner loop configurations are evaluated according to an approximate short-range potential $A_S(\boldsymbol{\alpha}(\mathbf{R}_m), \mathbf{r}_n)$, where the Born radii are held at a previous or "latent" state. The relation to the true short-range potential can similarly be written in terms of a short-range error potential $\varepsilon_S(\boldsymbol{\alpha}(\mathbf{R}_m),\mathbf{r}_n)$:

$$A_S(\alpha(\mathbf{R}_n), \mathbf{r}_n) = A_S(\alpha(\mathbf{R}_m), \mathbf{r}_n) + \varepsilon_S(\alpha(\mathbf{R}_m), \mathbf{r}_n) \tag{12}$$

where the coordinate state is $\mathbf{r}_n$, and the latent Born radii, $\boldsymbol{\alpha}(\mathbf{R}_m)$ are calculated from a previous step. Likewise, the true long-range potential can be described in terms of long-range error potential:

$$A_L(\alpha(\mathbf{R}_n), \mathbf{r}_n) = A_L(\alpha(\mathbf{R}_m), \mathbf{r}_n) + \varepsilon_L(\alpha(\mathbf{R}_m), \mathbf{r}_n) \tag{13}$$

For simplicity, these energies can be expressed in terms of the state indices only:

$$\begin{aligned} A_S(n, n) &= A_S(m, n) + \varepsilon_S(m, n) \\ A_L(n, n) &= A_L(m, n) + \varepsilon_L(m, n) \\ \varepsilon(m, n) &= \varepsilon_S(m, n) + \varepsilon_L(m, n) \end{aligned} \tag{14}$$

Where, n is the index of the current coordinate state and m is the index of the Born radii held at a previous state. We can simply recast the decomposition as

$$\begin{aligned} A(n, n) = A_S(n, n) + A_L(n, n) &= A_S(m, n) + \varepsilon(m, n) + A_L(m, n) \\ &= A(m, n) + \varepsilon(m, n) \end{aligned} \tag{15}$$

where the index of the coordinate state is first argument in each of the functions, and the index of the Born radii state is the second argument. While the error potential described in eq 14 contains both long and short-range terms, the idea of the sampling protocols is to treat the all of error potential terms as long-range terms. Using this new decomposition, we can define two different sampling protocols:

(1) In both protocols, start with the configuration $\mathbf{R}_i$, generate a number N_I of inner loop steps, where each trial configuration $\mathbf{r}_k$ is generated using eq 6. The Born radii are held at a latent state i, such that the short-range acceptance criterion is the following:

$$\frac{\mathrm{acc}_S(k'|k)}{\mathrm{acc}_S(k|k')} = \exp[-\beta(A_S(i, k') - A_S(i, k))] \tag{16}$$

(2) Take the final configuration from the inner loop to be the trial configuration $\mathbf{r}_j$ for the outer loop and apply either of two acceptance criteria:

(A) With error correction

$$\frac{\mathrm{acc}_L(j|i)}{\mathrm{acc}_L(i|j)} = \exp[-\beta(A_L(i, j) + \varepsilon(i, j) - A_L(i, i))] \tag{17}$$

(B) Without error correction

$$\frac{\mathrm{acc}_L(j|i)}{\mathrm{acc}_L(i|j)} = \exp[-\beta(A_L(i, j) - A_L(i, i))] \tag{18}$$

Protocol A rigorously obeys detailed balance, while protocol B is an approximation introduced to improve computational efficiency. It should be noted that the Born radii are completely updated in every outer loop calculation, regardless of protocol. The ideal error potential term would be narrowly distributed about a mean of zero, so that the distribution generated by neglecting the term would be nearly equivalent to the true distribution. The effect of the modification will be discussed in detail in the results section.

As a control, a "standard" Monte Carlo trajectory, or protocol S, was also studied. For the standard Monte Carlo protocol, the same trial move set was used, including steric screening, but with the Born radii updated at every step, with no decomposition of potentials. For every step, the acceptance criterion is simply:

$$\frac{\mathrm{acc}(j|i)}{\mathrm{acc}(i|j)} = \exp[-\beta(A(j, j) - A(i, i))] \tag{19}$$

Estimation of the Time to Convergence and Improvement Ration. To estimate the optimal number of inner steps, we express the total processor time T to compute a trajectory as

$$T = N_{O,T} \langle \mathrm{d}t/\mathrm{d}N_O \rangle \tag{20}$$

where $\langle \mathrm{d}t/\mathrm{d}N_O \rangle$ is the expectation value of the time required to generate an outer step. This is not a fixed value, since the innermost sampling loop samples an arbitrary number of configurations until a sterically allowed configuration is obtained. $N_{O,T}$ is the total number of outer steps, which includes the both the nonequilibrated steps, n_O, and equilibrated steps, N_O. This can also be expressed as

$$T = N_{O,T}(t_L + N_I t_S) \tag{21}$$

where t_s is the average time required to generate a single (sterically allowed) trial coordinate and evaluate the short-range potential. The rate t_L is the time required to evaluate the long-range potential, which includes the long-range energies and the time required to update the Born radii. This quantity does not need to be averaged, since there is no dependence on the number of steric clashes. N_I is the number of inner steps that are set for the simulation. Since statistics can only be gathered on the equilibrated outer steps, we can express N_O in terms of the standard error:

$$N_O = \frac{\sigma^2}{\varepsilon^2} g(N_I) \tag{22}$$

where σ is the variance of the energy over the entire equilibrated portion of the trajectory, ε is the desired error in the estimate of the energy, and $g(N_I)$ is the correlation interval, or distance between uncorrelated snapshots. This quantity is measured from the simulation, and will vary with the number of inner steps for a given system with all other conditions held constant. It is closely related to other measures of quality of Monte Carlo trajectories, such as acceptance ratio, and a low correlation interval often corresponds to a high acceptance ratio.

Since the number of steps required to equilibrate depends strongly on the initial condition, we shall overestimate this quantity by assuming that $n_O = N_O$. This varies in practice from a few correlation intervals to less than half of the number of outer steps. As long as the equilibration time is proportional to the number of equilibrated steps, it will cancel out in the improvement ratio calculation. Using this assumption, the estimated CPU time required for a converged trajectory is

$$T = 2\frac{\sigma^2}{\varepsilon^2} g(N_I)(t_L + N_I t_S) \tag{23}$$

Table 1. Simulation Data for Model System[a]

N_I	$N_{O,T}$	$<dt/dN_{O,T}>$ (s)	$<E> - <E>_{STD}$ (RT)	σ	ε	N_O (all)
S	250000	6.02	0.00	5.65	0.37	2366928
A-1	95000	6.29	−0.19	5.88	0.76	421025
A-50	40000	12.02	−0.01	5.83	0.21	195235
A-100	25000	16.37	0.13	5.83	0.19	122849
A-200	15000	26.08	0.17	5.86	0.22	74035
A-300	10000	34.37	0.25	5.82	0.24	48860
A-400	6000	43.44	0.12	5.83	0.31	29354
A-500	5000	51.48	0.13	5.91	0.37	24195
B-1	95000	6.20	1.77	5.96	0.28	393587
B-50	40000	11.11	1.58	6.07	0.07	198311
B-100	25000	16.01	1.79	6.12	0.08	123416
B-200	15000	24.72	1.89	6.12	0.08	73904
B-300	10000	32.86	1.71	6.19	0.10	48913
B-400	6000	41.40	1.69	6.09	0.11	29440
B-500	5000	50.47	1.83	6.18	0.12	24568

[a] Data shown summarizes the results for 10 simulations of each protocol and inner step setting. For the leftmost column, N_I is the number of inner steps. S indicates a standard protocol (no inner steps). For the remaining columns, protocol and number of inner steps are given. (A-50 represents protocol A using 50 inner steps). $N_{O,T}$ is the total number of steps simulated, including nonequilibrated portions of the trajectory. $dt/dN_{O,T}$ is the average time to generate an outer step, as described in the text. $<E> - <E>_{STD}$ (RT) is the average equilibrium energy minus the standard measurement, σ and ε are the standard deviation and standard error of the equilibrated energies. The rightmost column is the total number of equilibrated steps (across all simulations at the designated setting) used for the calculation.

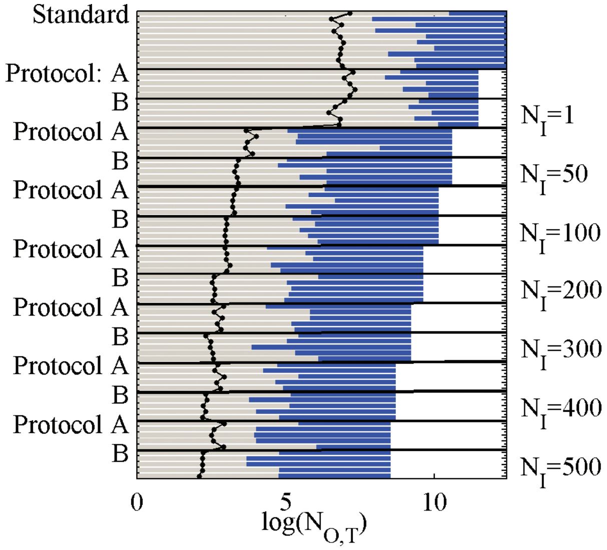


Figure 1. Summary statistics for validation data set. Bars represent the log of simulation lengths, and black dots connected with lines represent the correlation interval for that simulation. All simulations are run at 600 K. The blue portion of each bar is the unequilibrated portion, and the green portion is equilibrated. Different values are given for different runs, which are trajectories using the same settings, including initial condition, but assigned different random seeds. The natural log of the number of total steps, $N_{O,T}$, appears on the x-axis.

where the number of inner steps can be adjusted to locate the optimal computing time. As a measure of sampling efficiency, the following quantity can be expressed:

$$I = T_S/T \tag{24}$$

where I is the improvement, and T_s is the time required for a converged trajectory in a standard Monte Carlo protocol.

Convergence Determination and Error Estimation. Determination of the number of steps required for equilibration and the correlation interval was performed iteratively. Initially, the number of steps required for equilibration was estimated very approximately as 3000 for the standard trajectory, 1000 for $N_I = 1$, 50, 100, 200, 300, and 400 for the remaining inner step settings. To estimate the correlation time, an autocorrelation function of the energy was computed, and the correlation interval g was identified as the first place that the autocorrelation function crosses zero. This initial estimate is expected to overestimate the true correlation time since the trajectory may include nonequilibrated regions, which contain slow fluctuations toward the equilibrium state that would not be present in the stationary distribution. Using this initial estimate, a blocksize was assigned to have a value of g. A block standard deviation σ_B is computed at each point (using the points preceding the point of interest), and the trajectory was deemed to be converged if the block standard deviation was less than a nominal value $\sigma_B = 15k_BT$.

With this new estimate of the equilibrated region of the trajectory, another estimate of the correlation time was applied. To improve the estimate, the autocorrelation function was fit to a simple exponential $\exp(-\tau/\tau_D)$ where τ_D is the decay constant, or correlation time. For this procedure, a least-squares fit was performed where the sum of the squares of the errors between the function and the data points are weighted according to the inverse of error at that point. The error in the autocorrelation function is given by[45]

$$\varepsilon[C(\tau)] = \sqrt{\frac{g}{N_O - \tau}} \tag{25}$$

where $g = 1 + 2\tau_D$ is the correlation interval, or the number of steps between uncorrelated snapshots. Once a correlation time is obtained, the reverse cumulative averaging (RCA) method was used to obtain a better estimate of the location of the equilibrated region,[47] with the blocksize set to g. A confidence level of 85% was used to reject the hypothesis that the block averaged samples came from a normal distribution, according to the Shapiro−Wilk Test.[48,49] The location of the equilibrated portion of the trajectory depends heavily on the value of the blocksize, and vice versa, so 30 iterations of the blocksize and RCA convergence calculation were run. See Figure 1 for the convergence times, correlation

Table 2. Binding Pockets Studied[a]

label	protein	R_B	R_A	L_A	no. residues	$<N_C>$
A	db3 Antibody	1dba	1dbb	progesterone	30	0.54
B	thermolysin	1kr6	1kjo	Z-D glutamic acid	41	0.74
C	estrogen receptor	1err	3ert	raloxifene	73	0.65
D	PPAR-γ	1fm9	2prg	GI262570	65	0.75
E	PI3 kinase	2chx	2chw	PIK-039	45	0.65
F	CDK2	1buh	1dm2	hymenialdisine	46	0.73

[a] R_B is the receptor used in the simulation (without ligand), and R_A is a reference receptor with L_A bound to it. $<N_C>$ is defined as the total number of steric clashes divided by the number of sterically allowed steps.

intervals, and total simulation lengths for each simulation.

Preparation of Unbound receptors. The proteins studied are listed in Table 2. A few of the proteins had missing sidechains or loops, outside of the binding sites (>15Å) being studied. These were reconstructed in arbitrary configurations free of steric clashes using standard routines in the protein local optimization program. The sidechains to be sampled in the Monte Carlo were defined as those within 8 Å of any atom of the ligand in the holo structure. All calculations were performed in the absence of the ligand.

Composite Energy Histograms. In order to represent multiple simulations of the same sampling protocol as a single histogram, a superposition of individual energy histograms was computed. This is done to obtain better statistics so that detailed balance may be demonstrated for protocol A.

For each trajectory histogram, an error $\varepsilon_B = \sqrt(gn_B)$ was assigned at each bin point, where n_B is the number of entries in each bin. To generate the composite histograms for protocols A and B, each of the trajectory histograms for each protocol were superimposed with a weight proportional to the number of uncorrelated entries in each bin of each trajectory. The errors are computed a superposition of square of the errors of each trajectory, with the same weights used to compute the composite histograms. It should be noted that the sampling protocols produce the same distribution of energies, independent of number of inner steps chosen. The data from all ranges of inner steps can therefore be combined to form a single histogram. Since the error is computed using the autocorrelation times, the fact that the distributions fall within error suggest also that the correlation times are correctly estimated.

Timings. Since simulations were run on a variety of machines, smaller trajectories were collected to estimate the average time per outer step (see Table 1). Timings of the simulations were measured on a Linux machine, using a single CPU from a dual AMD Opteron CPU running at 2.2 GHz.

Results and Discussion

Comparison of Protocols Using Antibody DB3. To optimize the number of inner steps and other parameters of the algorithm, the binding pocket of apo antibody DB3 (1dba)[50,51] was selected as a model system. A total of three sampling protocols were explored, as defined in Theory and Methods. To compare the effect of neglecting the short-range error in the Born updates, identical simulations were run using protocols A (rigorous) and B (approximate). A single set of 10 trajectories using protocol S was also generated. The number of inner steps (N_I) was set to 1, 50, 100, 200, 300, 400, and 500. For each inner step setting, five trajectories were collected, starting from the same (nonequilibrium) initial condition with different random seeds. Since the backbone is held fixed, room temperature simulations tend to exhibit frustrated dynamics. To obtain better statistics, especially for protocol S, all simulations were run at 600 K. The goals of these simulations are twofold: (1) to generate sufficient statistics to demonstrate detailed balance and (2) to study the effect of adjusting the number of inner steps and protocol. A total of 80 separate trajectories were collected for the analysis. Figure 1 summarizes the pertinent information on these trajectories.

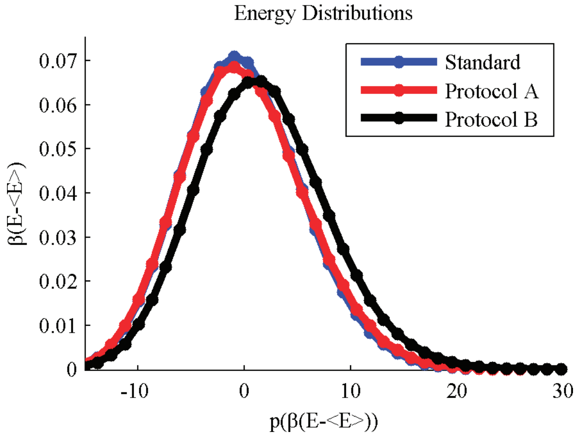


Figure 2. Protocol A distributions superimpose with standard energy histograms, and protocol B generates a similar approximate distribution. All simulations were run at 600 K, under the conditions summarized in Figure 1. Dimensionless energy is plotted on the *x*-axis, with the mean of the energies of the standard simulation $<E>$ subtracted from the energy (see Table 1). On the *y*-axis is the probability of observing that energy.

The average energies and standard errors of each simulation are in Table 2, and Figure 2 shows histograms of equilibrated energies for each sampling protocol. The energy distributions of protocols A and S (standard) appear to be equivalent. While error bars are not shown for clarity, the histograms superimpose to well within the estimated error. The energy distribution of protocol B is offset by roughly 1.75 RT, and is clearly from a different distribution than protocol A. The standard deviation of protocol B is larger by roughly 0.3 RT. The broader distribution and higher mean value is due to the more permissive approximation, which increases the number of states that are accepted.

The correlation interval is shown in Figure 3. A sharp decrease is observed from $N_I = 50-200$, which steadily

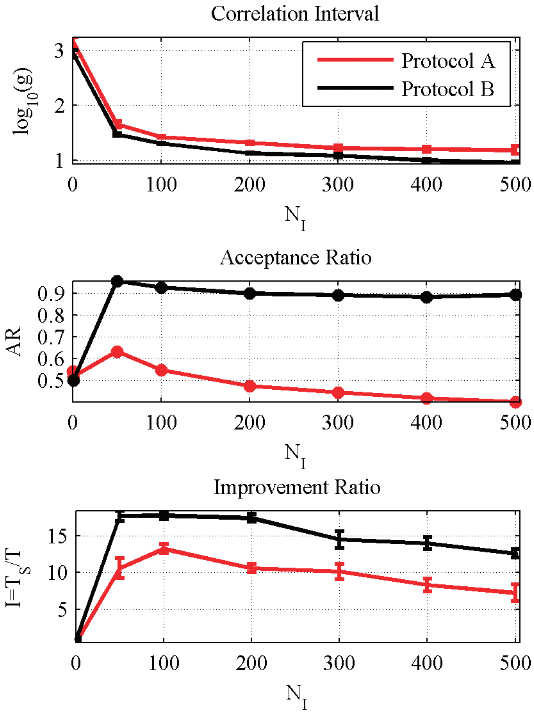


Figure 3. Approximate protocol provides slightly better performance, and optimal performance of both protocols is in the range of N_I = 50−200. (a) Logarithm of correlation interval. (b) Acceptance ratio. (c) Improvement ratio, as given by eq 24.

decreases over the remaining inner step settings. The acceptance ratio shows an initially sharp increase, since a smaller number of inner steps helps to generate better trial moves for the outer loop. As the number of inner steps increase however, the inner loop becomes less efficient at generating trial configurations. This effect is more prominent in protocol A, which is the rigorous approach. Figure 3c shows the relative improvement over protocol S (no inner steps). Optimal values are in the range N_I = 50−200. For both protocols A and B, a broad optimal range is observed, which suggests that this optimal range should hold for a wide variety of proteins.

Binding Pocket Studies. As a first application, we investigate the flexibility of sidechains in protein binding pockets. As a test set, we consider several proteins from Sherman et al.,[8] as well as PI3K.[52] The assumption of this work is that sidechains that show more flexibility in our ensembles will be capable of undergoing rearrangements upon binding ligands. Table 2 lists the binding pockets studied. For all trajectory data which is displayed, individual sidechains conformations were filtered such that no two conformations are less than an rmsd of 0.05 Å from one another.

Protocols A and B were used to generate sidechain ensembles, at a variety of temperatures. Temperatures >300 K were explored for three primary reasons. First, our goal is to predict conformational changes that could occur upon binding a ligand. In the limit of pure "conformational selection", the bound conformation of the protein would be populated significantly, or at least measurably, at ambient temperature. However, there can also be some additional conformational rearrangement of the 1protein to accommodate the ligand ("induced fit"), derived from the free energy of ligand binding. Here, we have essentially postulated that ligand binding can "induce" conformational changes that may not be observable with a room temperature thermal ensemble. It has been observed that sidechain rearrangements within binding pockets can be cost up to 4 kcal/mol of free energy.[15,16]

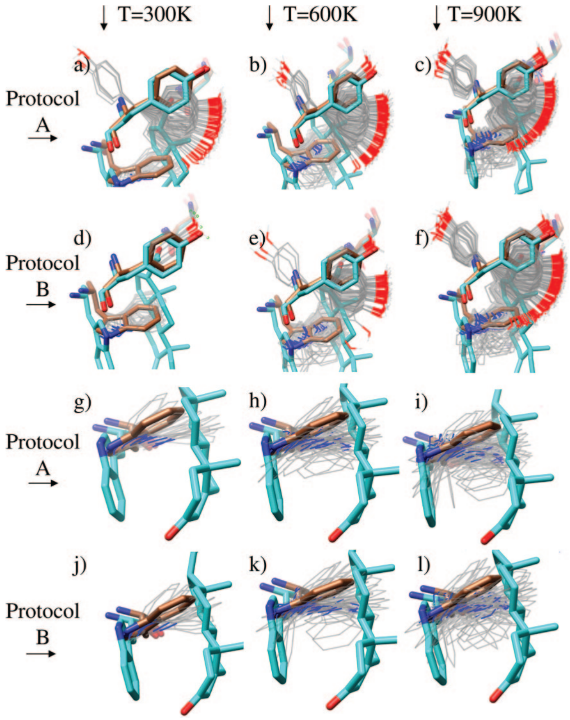


Figure 4. Distribution of sidechain configurations for Tyr97 and Trp100 of 1dba. Brown configurations are from the native structure, and cyan configurations are from the holo structure. Grey sidechains are distinct configurations from a sidechain trajectory at the given conditions. (a) Tyr97 at 300 K, protocol A. (b) Tyr97 at 600 K, protocol A. (c) Tyr97 at 900 K, protocol A. (d) Tyr97 at 300 K, protocol B. (e) Tyr97 at 600 K, protocol B. (f) Tyr97 at 900 K, protocol B. (g) Trp100 at 300 K, protocol A. (h) Trp100 at 600 K, protocol A. (i) Trp100 at 900 K, protocol A. (j) Trp100 at 300 K, protocol B. (k) Trp100 at 600 K, protocol B. (l) Trp100 at 900 K, protocol B.

Another reason for considering higher temperature distributions of 600 K is related to limitations of the energy function. In particular, it has been widely reported that generalized Born solvent models can overstabilize hydrogen bonds and salt bridge interactions.[39,53] This known limitation of the implicit solvent model will tend to result in reduced flexibility of charged residues at ambient temperatures.

Finally, the use of a rigid backbone will also reduce sidechain flexibility. The test cases were chosen in part because ligand binding does not induce large changes in backbone conformation; clearly, further algorithmic development, which will be reported in due course, is needed to deal with backbone fluctuations. When there is reason to believe that backbone changes are likely to be small, simply using a higher temperature may help to reduce artifacts due to the rigid backbone.

Ultimately, from the standpoint of identifying "flexible" sidechains in a binding site, we view the choice of temper-

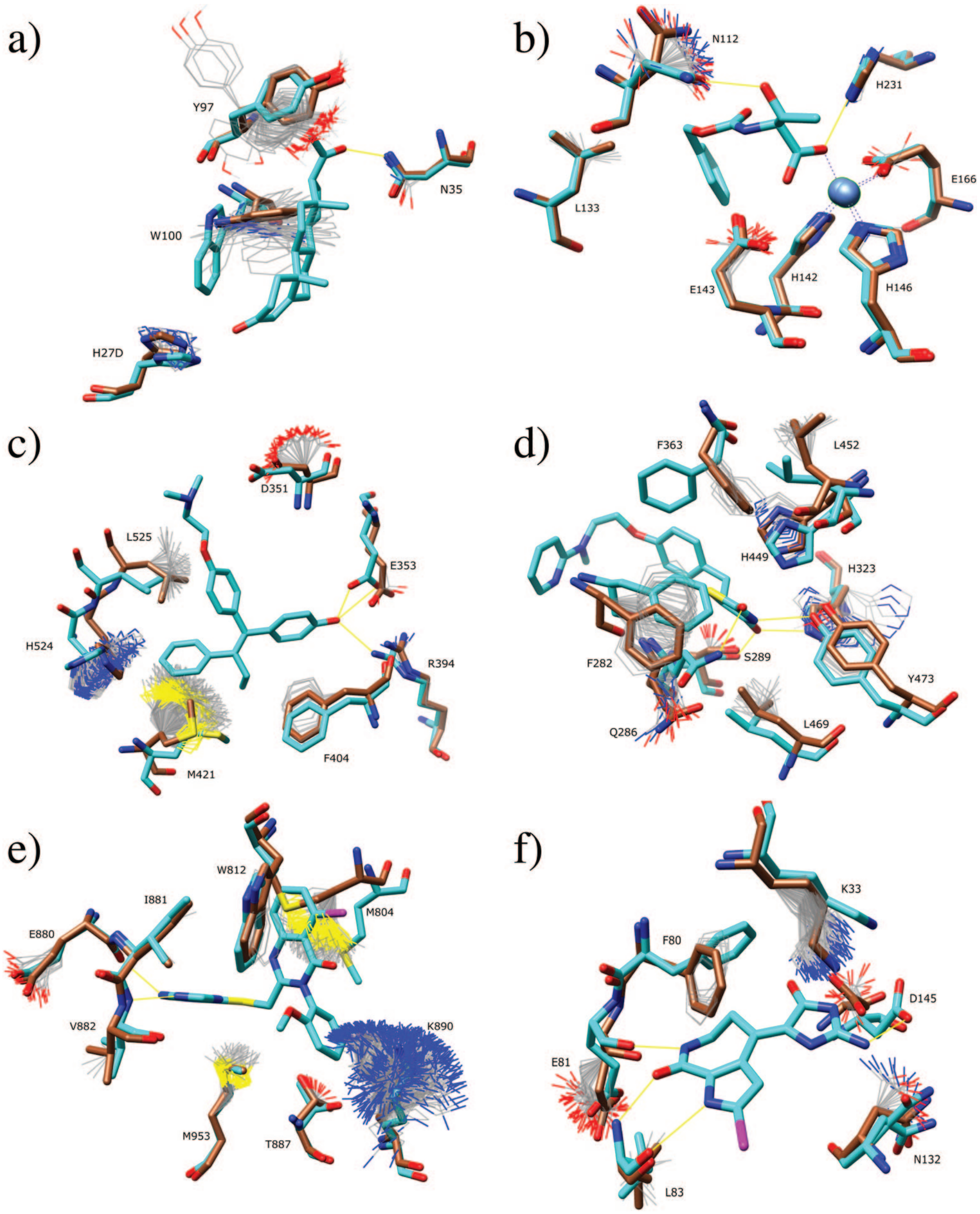


Figure 5. Binding pocket ensembles. Simulations are carried out in the absence of ligand at 600 K, with protocol B (no error correction). Ligand and bound (holo) structures are shown in cyan. Unbound native sidechains in starting configurations are shown in brown. The computed ensemble is shown as thin lines. The ligand from the holo structure is shown for reference. (a) DB3 antibody and progesterone. (b) Thermolysin and Z-D glutamic acid. (c) Estrogen receptor and raloxifene. (d) PPAR-γ and GI262570. (e) PI3 kinase and ligand PIK-039. (f) CDK2 and hymenialdisine.

ature as a user-definable parameter; in practice, performing simulations with multiple values of the temperature may be advisable. Note that, since the backbone is held fixed, the protein will not denature during the simulation, which provides considerable freedom in the choice of temperature and simulation protocol.

Antibody DB3.[50,51] For the DB3 antibody (Figures 4 and 5a), the primary conformational change between the two structures is the large movement of the Trp100 sidechain to accommodate 4-hydroxytamoxifen. We studied this system with both protocols A and B at $T = 300$, 600, and 900 K, with $N_I = 200$ (the upper end of the optimal range). It is encouraging to observe that the large conformational change in Trp100 is observed in the Monte Carlo simulations, performed without a ligand present, at 600 K using protocol B and at 900 K using protocol A. Two conformational states of Trp100 are observed: a low-population state where the sidechain is in a similar conformation as the holo structure and a high-population state where it is similar to the apo structure, although significant fluctuation is observed. Intermediate conformations are not observed suggesting a high energy barrier for the rotation.

The residues His27D and Asn35 show less flexibility in the simulations and also little conformational change between the apo and holo structures (Figure 5a). Tyr97, in contrast, appears to fluctuate in multiple basins. This is because it is mostly solvent exposed, and there is very little steric hindrance. The sidechain adopts similar conformations in the apo and holo structures. This does not necessarily imply a failure of the computational prediction, however. It is

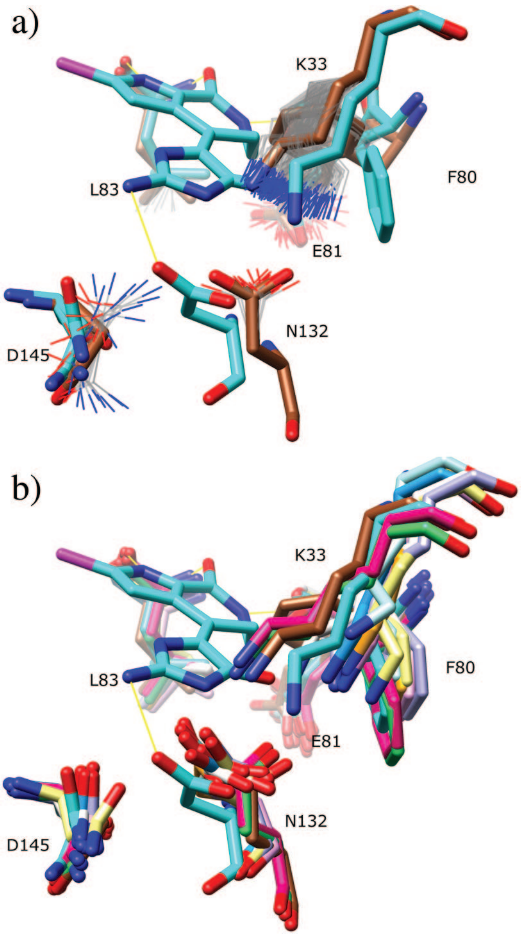


Figure 6. CDK2 salt bridge interaction. (a) Binding pocket ensemble and representation which is identical to Figure 5f, but from a different perspective. (b) Sidechains from CDK2 structures 1h24, 1h25, 1h26, 1h27, 1h28, 1hc1, 1pw2, 1w98, and 2jgz.

possible that this sidechain could adopt different conformations in complex with other ligands.

The magnitudes of fluctuations observed using protocols A and B for Trp100 and Tyr97 are similar (Figure 4). Since protocol B is slightly more efficient and appears to provide similar configurational diversity, it was used for the data presented for all the remaining binding pockets in Figure 5. In addition, we have chosen to use $T = 600$ K for the remainder of the test cases, because it provides a balance between sampling alternative conformations that may be important in ligand binding, but not so much diversity as to be uninformative. We reiterate that we view temperature as a user-adjustable parameter, and using multiple temperatures, as with this test case, may be advisable.

Thermolysin.[54] The residues His142, His146, and Glu66, which coordinate the Zn ion are correctly predicted to be rigid (Figure 5b). For this simulation, the zinc ion was included. The hydrogen bonding network of His231 is correctly preserved. Asn112 is predicted to be very flexible, and in fact rotates significantly upon ligand binding.

Estrogen Receptor.[55,56] Residues Leu525, Met421, and His524 all show significant flexibility in the simulations, and also undergo significant rearrangements upon binding 4-hydroxytamoxifin (Figure 5c). Glu353 and Arg394 display less flexibility due to the strong salt bridge. These show small conformational rearrangements upon binding the ligand due to formation of hydrogen bonds to it. Backbone rearrangements observed upon ligand binding, such as those seen in His524 and Leu525, are of course not captured by the sidechain MC simulations. As a rough guide, however, the ensemble correlates well with observed rearrangements.

PPAR-γ.[57] The hydrophobic residues Phe282, Leu452, and Leu469 display flexibilities that correspond to structural rearrangements upon ligand binding (Figure 5d). Phe363 fails to sample the bound configuration, and is the first of only two false negative cases from the entire data set (see CDK2). It is likely that this is due to the fact that the rigid backbone occupies a region which occludes the possibility of sampling an alternative state. His449 displays a narrow range of flexibility which corresponds to the displacement in the target structure. Tyr473 samples alternative solvent exposed configurations, similar to Tyr97 in the DB3 antibody. Gln286 displays flexibility and appears to sample some conformations similar to the holo conformation, to the extent that the slightly different backbone configurations permit.

PI3 Kinase.[52,58] All residues which do not undergo significant rearrangement upon ligand binding are predicted to be rigid in the simulations (Figure 5e). Glu880 and Lys890 display conformational diversity in the simulations which encompasses the observed apo and holo conformations. Met804 displays significant flexibility in the sidechain ensemble which encompasses the apo and holo conformations. The movement of this sidechain is critical for opening a hydrophobic pocket that is critical for ligand binding and specificity.

CDK2.[59–61] Residues Glu81, Leu83, and Asn132 each appear to display conformational diversity commensurate with the observed changes between the apo and holo structures (Figure 5f), while Phe80 is the second false negative of the data set. Lys33 displays flexibility, although it does not quite sample the bound configuration. Instead, in the absence of ligands, it forms a salt bridge with Asp145, which is disrupted by the hymenialdiside interaction in the bound form.

Figure 6a shows a closeup of the salt bridge which is transiently disrupted in the 600 K simulation. Figure 6b shows a superposition of multiple structures of CDK2 which display a similar structural diversity.

Conclusions and Future Directions

A novel application of the MTS-MC algorithm has been applied to sampling sidechain degrees of freedom in implicit solvent. Relative to a “simple” Monte Carlo algorithm without the use of inner steps, the multiscale approach increases the convergence by a factor of 10−15. Rapid steric screening provides an additional factor of 2−4 speed up, and other algorithmic details (rapid updates of energies when only a portion of the protein is moving) also contribute to efficiency. Applications to small molecule ligand binding sites in proteins demonstrate that the method can be used to efficiently sample large changes in sidechain conformations and identifies sidechains that may undergo conformational changes upon ligand binding.

Additional degrees of freedom can be incorporated into this approach in a straightforward manner. For example, local changes in backbone conformation can be included using analytical loop closure[62,63] methods with an appropriate Jacobian.[64] Such a method, which is under development, could be an efficient means of sampling conformational changes such as those that have been observed in the kinase DFG motif, or in loop latching as in TIM barrels,[65] in a way that obeys detailed balance and thus can capture entropy differences between states.

Acknowledgment. We thank Kevan Shokat (UCSF) for pointing out the PI3K system. All molecular figures were generated with UCSF Chimera.[66] This work was supported by NSF grant MCB-0346399, NIH grant GM081710, and the Sandler Program in Basic Science (UCSF). M.J. is a consultant to Schrödinger, Inc. Special thanks to Katarzyna Bernacki, Dave Mobley, Arjun Narayanan, and Chris McClendon for helpful discussions.

Note Added after ASAP Publication. This article was released ASAP on April 16, 2008 with minor errors in eqs 5, 33, 34, 38, and 39. The correct version was posted on April 25, 2008.

Appendix: Proof of Detailed Balance with a Short Range Cutoff

A more detailed accounting of the short- and long-range decompositions is presented. These details are omitted from the body of the text for clarity.

The use of a short- and long-range cutoff is a common way of improving calculation efficiencies. The advantage gained is in the infrequent updating of the long-distance interactions. To explicitly track the updating of the short- and long-range cutoffs, eqs 14 and 15 can be re-expressed as follows:

$$A(m,n) = S(l)A(m,n) + (1 - S(l))A(m,n) \tag{26}$$

Where $S(l)$ is a "switching function" of the coordinate state l, which divides the space over which the potential $A(m, n)$, as expressed in eq 15, is the potential at Born state m and coordinate state n. When the Born radii are evaluated based on the current coordinate state, the short- and long-range potentials can be expressed in terms of the current coordinate (and Born radii) state n, and latent cutoff state l:

$$\begin{aligned} A_S(l,m,n) &= S(l)A(m,n) \\ A_L(l,m,n) &= (1 - S(l))A(m,n) \end{aligned} \tag{27}$$

Since $S(l)$ is a function of the complete set of coordinates, a full update of the distances must be computed. The idea behind the use of the cutoff is to limit the number of times the full distance matrix is computed, as well as the full potential.

To this end, an efficient Monte Carlo protocol will update the switching function infrequently, while maintaining detailed balance or very nearly doing so. For the updating scheme that is used for the present work, detailed balance is rigorously maintained with regard to the short and long-range evaluations. The simplest form that the switching function can take is a simple distance cutoff, but more complicated forms, such as cell neighbor lists and other types of additive decompositions can be used. For this work, atoms are treated as short-range if any single atom within a sidechain is within a cutoff distance of another sidechain. Default settings that were developed for an optimal minimization strategy were used.[44] The cutoffs vary according to type of interaction. Each sidechain is identified as either charged or nonpolar. All atoms in the given sidechain are labeled as such. For nonpolar atoms interacting with nonpolar atoms, the cutoff is 15 Å. For charged−nonpolar interactions, the cutoff is 20 Å, and for charged−charged interactions, the cutoff is 30 Å. The updating scheme used for the current work is to update the switching function at the beginning of the each outer iteration of the sampling loop.

While the proof of detailed balance for the switching function updating scheme is independent of the Born radii updating scheme, the full bookkeeping of all latent states is presented here for completeness. Re-expressing the short- and long-range potentials in eq 13 with the short-range state made explicit gives the following:

$$\begin{aligned} A_S(l,m,n) &= A_S(l,n,n) - \varepsilon_S(l,m,n) \\ A_L(l,m,n) &= A_L(l,n,n) + \varepsilon_L(l,m,n) \\ \varepsilon(m,n) &= \varepsilon_S(l,m,n) + \varepsilon_L(l,m,n) \end{aligned} \tag{28}$$

The resulting (unnormalized) probability distributions are

$$\begin{aligned} p_S(l,m,n) &= e^{-\beta A_S(l,m,n)} \\ p_L(l,m,n) &= e^{-\beta A_L(l,m,n)} \\ p_\varepsilon(m,n) &= e^{-\beta \varepsilon(m,n)} \end{aligned} \tag{29}$$

where q is given by eq 5. Expressing the probability of a single state in terms of the decomposed states gives the following:

$$\begin{gathered} p(n) = e^{-\beta A(n,n)}/q \\ p(n) = p_S(l,n,n)p_L(l,n,n) = p_S(l,m,n)p_\varepsilon(m,n)p_L(l,m,n) \end{gathered} \tag{30}$$

Following the derivations presented in refs 35 and 43, the required detailed balance condition is

$$\begin{gathered} p(i)T(j|i) = p(j)T(i|j) \\ p_S(i,i,i)p_L(i,i,i)T(j|i) = p_S(j,j,j)p_L(j,j,j)T(i|j) \end{gathered} \tag{31}$$

where $T(j|i)$ is the probability of transitioning from coordinate state i to j. Expanding this expression gives:

$$\begin{aligned} &p_S(i,i,i)p_L(i,i,i)\alpha(j|i)\mathrm{acc}_L(j|i) \\ &\qquad = p_S(j,j,j)p_L(j,j,j)\alpha(i|j)\mathrm{acc}_L(i|j) \end{aligned} \tag{32}$$

where $\alpha(j|i)$ and $\mathrm{acc}_L(j|i)$ are the selection and acceptance probabilities of outer state j from state i. Following the MTS-MC derivation,[35] the probability of selecting state j from state i is given by the following:

$$\alpha(j|i) = T_S^{(N_I)}(j|i) \tag{33}$$

where the above transition probability is the product of the individual transition probabilities of the inner loop

$$T_S^{(N_I)}(j|i) = T_S(1|i)\left[\prod_{k=1}^{N_I-2} T_S(k+1|k)\right]T_S(j|N_I - 1) \tag{34}$$

In the short-range, or inner loop of sampling, neither the switching function nor the Born radii are updated, so that each step obeys the following detailed balance relation:

$$p_{\mathrm{S}}(i,i,k)T_{\mathrm{S}}(k'|k) = p_{\mathrm{S}}(i,i,k')T_{\mathrm{S}}(k|k') \tag{35}$$

The transition between outer states j and i obeys the following detailed balance relation:

$$p_{\mathrm{S}}(i,i,i)T_{\mathrm{S}}^{(N_{\mathrm{I}})}(j|i) = p_{\mathrm{S}}(i,i,j)T_{\mathrm{S}}^{(N_{\mathrm{I}})}(i|j) \tag{36}$$

Combining eqs 32−35 and solving for the ratio of acceptance probabilities gives the following:

$$\left.\frac{\mathrm{acc}_{\mathrm{L}}(j|i)}{\mathrm{acc}_{\mathrm{L}}(i|j)}\right|_{\mathrm{TRUE}} = \frac{p_{\mathrm{L}}(j,j,j)p_{\mathrm{S}}(j,j,j)}{p_{\mathrm{L}}(i,i,i)p_{\mathrm{S}}(i,i,j)} \tag{37}$$

Protocols A and B follow the same updating scheme for the switching functions. The acceptance probability for protocol A is expressed in eq 17 as follows:

$$\left.\frac{\mathrm{acc}_{\mathrm{L}}(j|i)}{\mathrm{acc}_{\mathrm{L}}(i|j)}\right|_{\mathrm{A}} = \frac{p_{\mathrm{L}}(i,i,j)p_{\varepsilon}(i,j)}{p_{\mathrm{L}}(i,i,i)} \tag{38}$$

The ratio of eqs 37 and 38 is unity

$$\begin{aligned}&\left(\left.\frac{\mathrm{acc}(j|i)}{\mathrm{acc}(i|j)}\right|_{\mathrm{TRUE}}\right/\left(\left.\frac{\mathrm{acc}(j|i)}{\mathrm{acc}(i|j)}\right|_{\mathrm{A}}\right.\\ &= \frac{p_{\mathrm{L}}(j,j,j)p_{\mathrm{S}}(j,j,j)}{p_{\mathrm{L}}(i,i,i)p_{\mathrm{S}}(i,i,j)}\cdot\frac{p_{\mathrm{L}}(i,i,i)}{p_{\varepsilon}(i,j)p_{\mathrm{L}}(i,i,j)}\\ &= \frac{p_{\mathrm{L}}(j,j,j)p_{\mathrm{S}}(j,j,j)}{p_{\mathrm{S}}(i,i,j)p_{\varepsilon}(i,j)p_{\mathrm{L}}(i,i,j)} = \frac{p(j)}{p(j)} = 1\end{aligned} \tag{39}$$

and therefore, the sampling scheme described by eqs 37 and 38 rigorously obeys detailed balance. For all equations in the body of the text, the state of the switching function is not shown, but is updated according to the scheme described. It should be noted, however that the "standard" protocol is not updated according to this scheme, since there is no need to express the energies in terms of the latent states.

The acceptance probabilities for protocol B, as given in 18, are as follows:

$$\left.\frac{\mathrm{acc}_{\mathrm{L}}(j|i)}{\mathrm{acc}_{\mathrm{L}}(i|j)}\right|_{\mathrm{B}} = \frac{p_{\mathrm{L}}(i,i,j)}{p_{\mathrm{L}}(i,i,i)} \tag{40}$$

The ratio of the true acceptance probabilities is equivalent to the acceptance probabilities given in eq 37, and the ratio is given simply as follows:

$$\begin{aligned}\left(\left.\frac{\mathrm{acc}(j|i)}{\mathrm{acc}(j|i)}\right|_{\mathrm{A}}\right/\left(\left.\frac{\mathrm{acc}(j|i)}{\mathrm{acc}(j|i)}\right|_{\mathrm{B}}\right. &= \frac{p_{\mathrm{L}}(i,i,j)p_{\varepsilon}(i,j)}{p_{\mathrm{L}}(i,i,i)}\,\frac{p_{\mathrm{L}}(i,i,i)}{p_{\mathrm{L}}(i,i,j)}\\ &= p_{\varepsilon}(i,j) = \exp[-\beta\varepsilon(i,j)]\end{aligned} \tag{41}$$

References

(1) Dunbrack, R. L., Jr.; Karplus, M. *Nat. Struct. Biol.* **1994**, *1*, 334.

(2) Dunbrack, R. L., Jr.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661.

(3) Xiang, Z.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421.

(4) Kuhlman, B.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10383.

(5) Jiang, L.; Kuhlman, B.; Kortemme, T.; Baker, D. *Proteins* **2005**, *58*, 893.

(6) Fiser, A.; Do, R. K.; Sali, A. *Protein Sci.* **2000**, *9*, 1753.

(7) Fiser, A.; Sali, A. *Methods Enzymol.* **2003**, *374*, 461.

(8) Sherman, W.; Day, T. J.; Jacobson, M.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534.

(9) Meiller, J.; Baker, D. *Proteins* **2006**, *65*, 538.

(10) Ferrari, A. M.; Wei, B.; Constantino, L.; Shoichet, B. K. *J. Med. Chem.* **2004**, *47*, 5076.

(11) Voigt, C. A.; Gordon, D. B.; Mayo, S. L. *J. Mol. Biol.* **2000**, *299*, 789.

(12) Jain, T.; Cerutti, D. S.; McCammon, J. A. *Protein Sci.* **2006**, *15*, 2029.

(13) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1982**, *161*, 269.

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

(15) Mobley, D. *J. Chem. Theory Comput.* **2007**, *3*, 1231.

(16) Mobley, D.; Graves, A.; Chodera, J. D.; McReynolds, A.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118.

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

(18) Rosenbluth, M. N.; Rosenbluth, A. W. *J. Chem. Phys.* **1955**, *23*, 356.

(19) Deem, M. W. *J. Chem. Phys.* **1999**, *111*, 6625.

(20) Deem, M. W. *Mol. Phys.* **1999**, *97*, 559.

(21) Dinner, A. R. *J. Comput. Chem.* **2000**, *21*, 1132.

(22) Ulmschneider, J. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2004**, *126*, 1849.

(23) Li, Z. Q.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611.

(24) Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. *J. Mol. Biol.* **2002**, *317*, 493.

(25) 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.

(26) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins* **2002**, *48*, 404.

(27) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B.* **1998**, *102*, 10983.

(28) W. Clark Still, A. T.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(29) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517.

(30) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B.* **2001**, *105*, 6474.

(31) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351.

(32) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597.

(33) Tuckerman, M.; Berne, B. J. *J. Chem. Phys.* **1990**, *94*, 1465.

(34) Tuckerman, M.; Berne, B. J.; G.J., M. *J. Chem. Phys.* **1992**, *97*, 1990.

(35) Hetenyi, B.; Bernacki, K.; Berne, B. J. *J. Chem. Phys.* **2002**, *117*, 8203.

(36) Michel, J.; Taylor, R.; Essex, J. *J. Chem. Theory Comput.* **2006**, *2*, 732.

(37) Roux, B. *Implicit Solvent Models*; Marcel Dekker: New York, 2001.

(38) Yu, Z.; Jacobson, M. P.; Friesner, R. A. *J. Comput. Chem.* **2005**, *27*, 72.

(39) Jacobson, M. *J. Phys. Chem. B* **2004**, *108*, 6643.

(40) Verlet, L. *Phys. Rev.* **1968**, *165*, 201.

(41) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *55*, 5422.

(42) Bernacki, K.; Hetenyi, B.; Berne, B. J. *J. Chem. Phys.* **2004**, *121*, 44.

(43) Gelb, L. D. *J. Chem. Phys.* **2003**, *118*, 7747.

(44) Zhu, K. FriesnerR. A.; JacobsonM. P. *J. Comput. Chem.* **2006**.

(45) D Frenkel, B. S. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: Boston, 2002.

(46) Chen, B.; Siepmann, J. I. *Theor. Chem. Acc.* **1999**, *103*, 87.

(47) Wei, Yang; Martin Karplus, R. B.-P. *J. Chem. Phys.* **2004**, *120*.

(48) Shapiro, M. B. W.; Chen., H. J. **1968**, *63*, 1343.

(49) Shapiro, S.; Wilk, M. B. *Biometrika* **1965**, *52*, 591.

(50) Arevalo, J. H.; Hassig, C. A.; Stura, E. A.; Sims, M. J.; Taussig, M. J.; Wilson, I. A. *J. Mol. Biol.* **1994**, *241*, 663.

(51) Arevalo, J. H.; Stura, E. A.; Taussig, M. J.; Wilson, I. A. *J. Mol. Biol.* **1993**, *231*, 103.

(52) Knight, Z. A.; Gonzalez, B.; Feldman, M. E.; Zunder, E. R.; Goldenberg, D. D.; Williams, O.; Loewith, R.; Stokoe, D.; Balla, A.; Toth, B.; Balla, T.; Weiss, W. A.; Williams, R. L.; Shokat, K. M. *Cell* **2006**, *125*, 733.

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

(54) Senda, M.; Senda, T.; Ogi, T.; Kidokoro, S.; Stihle, R.; Boroni, E.; Hennig, M. *Acta Crystallogr.* **2002**, *58*, C278.

(55) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. *Cell* **1998**, *95*, 927.

(56) Pike, A. C.; Brzozowski, A. M.; Walton, J.; Hubbard, R. E.; Bonn, T.; Gustafsson, J. A.; Carlquist, M. *Biochem. Soc. Trans.* **2000**, *28*, 396.

(57) Gampe, R. T., Jr.; Montana, V. G.; Lambert, M. H.; Miller, A. B.; Bledsoe, R. K.; Milburn, M. V.; Kliewer, S. A.; Willson, T. M.; Xu, H. E. *Mol. Cell* **2000**, *5*, 545.

(58) Nolte, R. T.; Wisely, G. B.; Westin, S.; Cobb, J. E.; Lambert, M. H.; Kurokawa, R.; Rosenfeld, M. G.; Willson, T. M.; Glass, C. K.; Milburn, M. V. *Nature* **1998**, *395*, 137.

(59) Meijer, L.; Thunnissen, A. M.; White, A. W.; Garnier, M.; Nikolic, M.; Tsai, L. H.; Walter, J.; Cleverley, K. E.; Salinas, P. C.; Wu, Y. Z.; Biernat, J.; Mandelkow, E. M.; Kim, S. H.; Pettit, G. R. *Chem. Biol.* **2000**, *7*, 51.

(60) Bourne, Y.; Watson, M. H.; Hickey, M. J.; Holmes, W.; Rocque, W.; Reed, S. I.; Tainer, J. A. *Cell* **1996**, *84*, 863.

(61) Groban, E. S.; Narayanan, A.; Jacobson, M. P. *PLoS Comput. Biol.* **2006**, *2*, e32.

(62) Coutsias, E. A.; Seok, C. L.; Jacobson, M. P.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 510.

(63) Go, N.; Scheraga, H. A. *Macromolecules* **1969**, *3*, 178.

(64) Dodd, L. R.; Boone, T. D.; Theodorou, D. N. *Mol. Phys.* **1993**, *78*, 961.

(65) Wong, S.; Jacobson, M. P. *Proteins* **2008**, *71*, 153.

(66) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605.

CT700334A

JCTC Journal of Chemical Theory and Computation

Computational Studies of the X-Linked Inhibitor of Apoptosis Complex Formation with Caspase-9 and a Small Antagonist

George A. Kaminski*

Central Michigan University, Mount Pleasant, Michigan 48859

Received January 17, 2008

Abstract: Apoptosis is self-programmed cell death. The X-linked inhibitor of apoptosis (XIAP) is known to inhibit caspase proteins, the key players in apoptosis. When this happens, the cells become cancerous as they cannot die naturally. XIAP inhibitors are often overexpressed in cancer tissue. Presented in this article are the results of simulations of XIAP-caspase and XIAP-antagonist complexes. It has been previously established experimentally that low intensity ultrasound promotes apoptosis and increases the therapeutic effect of some XIAP-caspase interaction antagonists. The resulting calculated complex formation energies produced in this work were used with a simple multiscale model as an example of applying such energetic results for estimating the effects of ultrasound on these complexes. The microscopic simulations have been carried out with molecular mechanics employing an all-atom description of the molecules with the OPLS-AA and polarizable force field (PFF) formalisms. It has been determined that the interaction energies in the XIAP-caspase-9 pair with both OPLS and PFF are roughly the same and in the 30–40 kcal/mol range, while PFF predicts a higher magnitude of energy of the XIAP-antagonist complex formation (ca. 100 kcal/mol vs ca. 40 kcal/mol), thus probably being more adequate in reproducing the inhibition abilities of this low molecular weight antagonist. The presented study of the ultrasound effect leads to the conclusion that it is most likely based on the cavitation accompanying the ultrasound irradiation of the cells and not on a simple frequency resonance, as was suggested by some authors.

I. Introduction

A. Apoptosis As a Natural Anticancer Mechanism, Caspase Proteins, X-Linked Inhibitor of Apoptosis. Apoptosis is a specific type of self-induced cell death.[1] Alteration of apoptosis pathways can lead to its being underpresent, which causes the cells to become "immortal" (cancer).[1b] This is why recently there has been much attention directed toward understanding apoptosis and its reactivation after inhibition, especially in the area of cancer research.[2]

Apoptosis is executed by caspases, a family of cystein proteases. The critical involvement of caspases in apoptosis has been documented and discussed in a number of works.[1,2] Clearly, inhibition of caspases leads to a failure of apoptosis. The Inhibitors of Apoptosis (IAP) are a family of proteins which strongly interact, bind, and inhibit caspases. XIAP (or X-linked inhibitor of apoptosis) is one of such molecules receiving much attention recently.[3] Each XIAP protein contains copies of the 80 residue baculoviral IAP repeat (BIR). Each BIR domain has a distinct function. For example, BIR3 of XIAP efficiently inhibits caspase-9 protein. The BIR3 domain of XIAP captures caspase-9 in its inactive conformation and prevents activation.[1b]

Thus, the XIAP effectively stops apoptosis. Disruption of the XIPA (BIR3) – caspase-9 complex (inhibition of XIAP) is viewed as a way to induce apoptosis in cancer cells and to work as an antitumor agent. A number of experimental studies of possible pro-apoptosis agents, including Smac/DIABLO, peptides similar to Smac/DIABLO, and small molecules, have been undertaken.[4] Promoting apoptosis is a valid and actively pursued target in current cancer research.

* Corresponding author e-mail: kamin1ga@cmich.edu.

10.1021/ct8000188 CCC: $40.75 © 2008 American Chemical Society
Published on Web 03/29/2008

This article is intended to make a molecular modeling contribution in the area of studying apoptosis inhibition.

B. Simulations of Intermolecular Interactions – Fixed Charges and Polarizable Force Fields. One of the most common approaches in simulation of proteins is the fixed-charges model. In this case, the electrostatic interactions between particles are represented by attraction or repulsion between constant predetermined Coulomb charges assigned to the atomic sites. This scheme is implemented in such widely used force fields as AMBER,[14] MMFF,[15] and OPLS-AA.[16] While this approach usually predicts the structure of biological systems and the relative binding energies rather well, and the computational efficiency is high, there are certain limitations to the technique. Since the Coulomb charges on atoms do not change in the course of simulations, there is no way the simulated electrostatics can adequately respond to a changing electrostatic environment. For example, when a molecule is immersed in a polar solvent, such as water, it usually becomes more polar (a water molecule in gas phase has a dipole moment of ca. 1 Debye lower than the same molecule in pure liquid water[17]). If a force field utilizing permanent fixed charges is parametrized to reproduce liquid-state properties, it will inevitably be overpolarized for gas-phase or other low dielectric constant media. This is why absolute binding energies in gas phase can be overestimated by fixed-charges force fields by as much as 2.5 kcal/mol even for small molecules.[18] It is therefore highly desirable to have a polarizable electrostatic model, which can readjust as the environment changes. In this case, magnitudes of the charges can change or point electrostatic dipoles are induced. A number of polarizable force fields have emerged in recent years, with the scope of the applications ranging from neat liquids to protein–ligand complexes in solutions and computing absolute acidity constants.[19,20]

Since the purpose of the presented work is directly dependent upon estimating binding energies, it is natural to assume that the explicit treatment of electrostatic polarization could be inportant. This is why this work was done with using both fixed-charges OPLS-AA force field[16,21] and a recently developed polarizable force field for proteins and small molecules.[18,19e] Comparison of the results is aimed at identifying the situations in which employing polarizable force fields is critical.

C. Application of Low-Intensity Ultrasound As an Apoptosis-Promoting Technique. Uses of ultrasound in medical applications are numerous and range from imaging in diagnostics to ultrasound treatments of tumors.[5–10] Effects of low-intensity ultrasound are less well studied than those of the high-intensity one, but they are nevertheless quite promising, especially in the area of cancer research. For example, death rate of human ovarian carcinoma cells have been reported to increase as a result of ultrasound sonification, while the energy directly associated with the ultrasound itself was clearly not sufficient to kill them.[11] Apoptosis in human leukemic cells can be induced by low-energy ultrasound, which suggests new ways of anticancer therapy.[10] Enhancement of chemotherapy by sonification has also been reported. Exposure to ultrasound enhances cytotoxicity of anticancer chemicals to cancer cells. As a result, the dosage of a drug can be reduced and a patient's tolerance to chemotherapy improved.[7] Ultrasound can synergize the effects of adriamycin, cisplatin, 5-fluoraurcail, arabinosyl cytosine, boron hydrochloride monohydrate, diazoquononem, and 4′-O-tetrahydropyranyladriamicyn.[7] The synergy has been confirmed in ovarian cancer, breast cancer, cervical cancer, and leukemia.

While the apoptosis inducing effect of ultrasound sonification has been established experimentally, the exact mechanism of this process still remains to be understood. Currently, three hypotheses have been put forward.[7] The first one suggests that the ultrasound irradiation causes conformational shifts, such as, for example, turning an inactive form of caspase proteins into active ones. The second hypothesis is the resonant frequency one, which states that the frequency of the irradiating ultrasound is close to the frequency of, for example, XIAP-caspase bonding. Thus, the ultrasound directly destroys the harmful caspase inhibition complex in a targeted manner. Finally, the third hypothesis is that the destruction of the cancer cells is caused by cavitation. Cavitation is a phenomenon which accompanies ultrasound propagation in liquids, such as water. An ultrasound wave creates zones of increased and decreased pressure. When the pressure is sufficiently decreased, a gas bubble emerges. Then the pressure increases again, and the bubble collapses. This cycle is repeated with the frequency of the ultrasound. The collapse of the air bubbles creates shock waves, and the pressure in these waves can reach as high as 40–60 kbar.[12,13] Therefore, the pressure range in the medium can greatly exceed the nominal ultrasound wave pressure amplitude of ca. 1.5 kbar.

The work presented in this manuscript explores the frequency resonance and cavitation action hypotheses at the microscopic level to cast light on the mechanism of the low-intensity ultrasound-induced apoptosis in cancer cells.

The remainder of the paper is organized as follows. Section II describes the methodology involved in computing microscopic protein–ligand interaction energies and the mechanistic model used to assess the effect of ultrasound upon the complex formation. Section III presents results of the calculations. These are followed by conclusions in Section IV.

II. Methods

A. Calculating Intermolecular Interaction Energies. First, interaction energies of the BIR3 domain of XIAP with the caspase-9 protein and a small molecular antagonist were calculated using both OPLS-AA and polarizable force field (PFF). The initial geometries of the complexes were taken from Protein Data Bank structures 1NW9 and 1TFQ, respectively. Hydrogen atoms were added to the structures using the Maestro program.[22] Then each complex was truncated so that only those residues with at least one atom within a cutoff distance of 7.5 Å of any atom of the other molecule in the complex were considered and all the other residues discarded. The small antagonist[23] shown in Figure 1 was not truncated in any way.

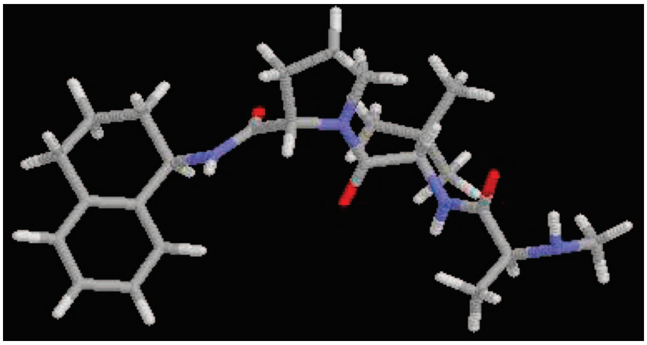


Figure 1. Antagonist to the XIAP-caspase-9 interaction.

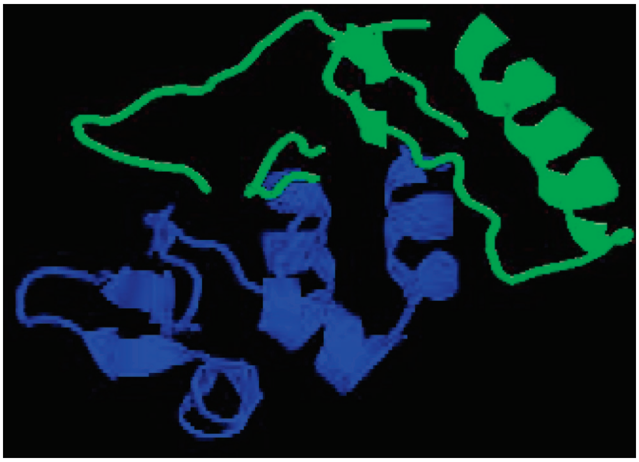


Figure 2. Explicitly modeled part of the XIAP-BIR3 interacting with caspase-9 protein. Residues present: chain A (blue): Leu256–Arg258, Val279–His346; chain B (green): Tyr153, Gln240–Gly248, Ala316–Gln320, Ser333–Thr347, Asp379–Lys410; Zn^{2+} ion (not shown).

Figures 2 and 3 demonstrate parts of the molecules included in simulations.

For each of the two complexes, a series of energy minimizations was performed. The parts of the complexes (chains A and B in the first case, chain A and the ligand in the second one) were moved with respect to each other prior to the optimizations in order to obtain interaction energy as a function of distance between the parts. The parts were displaced along the line passing through the centers of masses of the parts in their original full form (as in the PDB files, not truncated). Geometry of the ligand (red in Figure 3) was fixed. For all the protein parts, backbone geometry was fixed in the course of the energy minimizations, and the side chains were completely flexible. Interaction energies were computed using the OPLS-AA force field[16,21] and, in separate runs, the complete polarizable force field for proteins[19e] and polarizable force field for small molecules.[18] Geometry optimizations were performed in continuum PBF solvent corresponding to water for the polarizable force field[19e] and the standard IMPACT SGB model in the OPLS-AA calculations. IMPACT software suite was used for all the energy minimizations.[24] A conjugate gradient technique was employed with the convergence criterion for the final energy gradient set to 0.05 kcal/mol/Å.

After the dependence of the interaction energies on the distance between the parts of the complexes was obtained as described above, it was used in the mechanistic model introduced below to determine the effect of ultrasound irradiation on the XIAP-BIR3 complexes with caspase-9 and the small molecular antagonist.

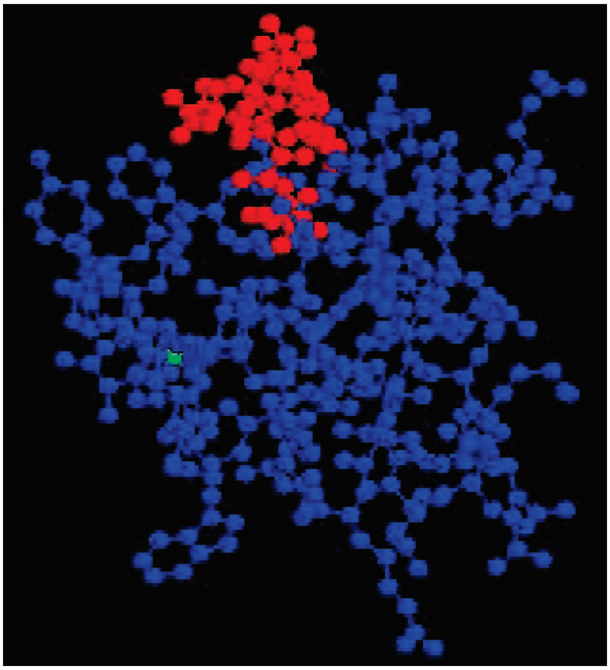


Figure 3. Explicitly modeled part of the XIAP-BIR3 interacting with the small antagonist. Residues present: chain A (blue): Tyr277–Tyr324; complete ligand (red); Zn^{2+} ion (green).

B. Force Fields. We have used both a polarizable force field (PFF) and a fixed-charges OPLS-AA. The procedure for building the PFF has been described elsewhere.[18] In the essence, the electrostatic interactions are represented by interactions of fixed charges and inducible point dipoles with each other. Fourier series were employed for the torsional energy and harmonic bond stretching and angle bending parameters were used.

In case of the fixed-charges OPLS force field, the key difference was that the nonbonded part was calculated as

$$E_{\mathrm{nb}} = \sum_{i<j} [q_i q_j e^2 / r_{ij} + 4\varepsilon_{ij}(\sigma_{ij}^{12} / r_{ij}^{12} - \sigma_{ij}^{6} / r_{ij}^{6})] f_{ij} \qquad (1)$$

The summation runs over all the pairs of atoms $i < j$ on molecules A and B or A and A for the intramolecular interactions. Moreover, in the latter case, the coefficient f_{ij} is equal to 0.0 for any *i-j* pairs 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 exactly 3 bonds) and $f_{ij} = 1.0$ for all the other cases. Standard OPLS-AA parameters were used.

C. Mechanistic Model for the Effect of Ultrasound Irradiation on the Protein–Ligand Complexes. The data from the intermolecular interactions simulations were used in a simple mechanistic model which has been devised to qualitatively estimate effects of ultrasound on the complexes in hand. While other mechanistic models had been proposed before,[25] none of them were combined with an explicit all-atom simulations to provide a detailed description of processes in a small protein–ligand complex. The

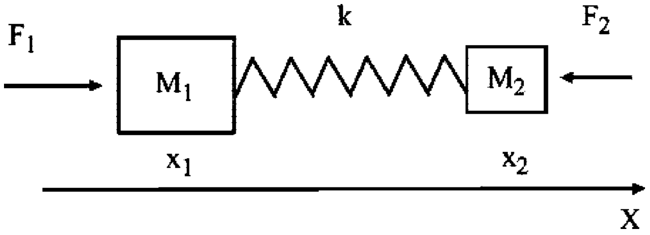


Figure 4. Mechanistic model for the protein–ligand complex.

model is somewhat crude, but it permits the drawing of conclusions in qualitative agreement with the available experimental data. The protein–ligand complex is schematically shown in Figure 4.

The interacting protein and ligand are represented as masses M_1 and M_2 located at one-dimensional positions x_1 and x_2 and connected by a spring with a constant k. In addition, the masses are experiencing external forces $F_1 = P{\bullet}A_1$ and $F_2 = P{\bullet}A_2$, where P is the external pressure and A_1 and A_2 are effective areas of the molecules. $A = \pi d^2/4$, where d is the largest distance between two atoms in a molecule. In the presence of ultrasound, $P = P_0 \cos(\omega t)$, where P_0 and ω are the amplitude and angular frequency of the ultrasound wave, and t stands for time. Then the equations of motion for the masses 1 and 2 are

$$\begin{aligned} M_1\ddot{x}_1 &= -k(x_1 - x_2) + F_1 \\ M_2\ddot{x}_2 &= -k(x_2 - x_1) - F_2 \end{aligned} \tag{2}$$

Dividing each equation by the corresponding mass and subtracting the second one from the first

$$\ddot{x}_1 - \ddot{x}_2 = \left(-\frac{k}{M_1} - \frac{k}{M_2}\right)(x_1 - x_2) + \frac{F_1}{M_1} + \frac{F_2}{M_2} \tag{3}$$

We are only interested in relative motion of the molecules, thus we can disregard the result of adding the eqs 2 describing the motion of the center of mass. Introducing a new variable $y = x_1 - x_2$

$$\ddot{y} = -\frac{\mathrm{k}(\mathrm{M}_1 + M_2)}{M_1 M_2} y + \left(\frac{F_1}{M_1} + \frac{F_2}{M_2}\right) \tag{4}$$

Using the reduced mass $\mu = M_1M_2/(M_1{+}M_2)$ and introducing $a(t) = a_0\cos(\omega t) = F_1/M_1 + F_2/M_2 = P{\bullet}A_1/M_1 + P{\bullet}A_2/M_2 = (A_1/M_1 + A_2/M_2){\bullet}P = (A_1/M_1 + A_2/M_2){\bullet}P_0 \cos(\omega t)$, thus $a_0 = P_0{\bullet}(A_1/M_1 + A_2/M_2)$, and, from eq 4

$$\ddot{y} + \frac{k}{\mu} y = a_0 \cos(\omega t) \tag{5}$$

or, with $k/\mu = \omega_0{}^2$

$$\ddot{y} + \omega_0^2 y = a_0 \cos(\omega t) \tag{6}$$

Equation 6 represents the classical problem of driven oscillatory motion. The solution (after a certain equilibration time) is

$$y = \frac{a_0}{\omega_0^2 - \omega^2} \cos\omega t) \tag{7}$$

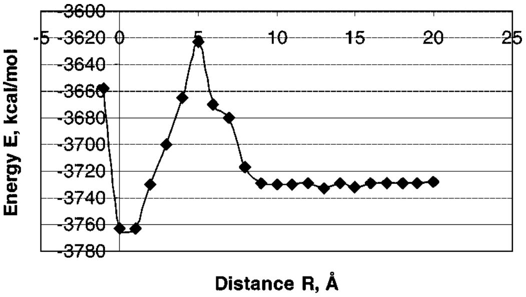


Figure 5. Energy *E* of XIAP-BIR3 interaction with caspase-9 as a function of the distance *R* between the molecules. Computed with the OPLS-AA force field.

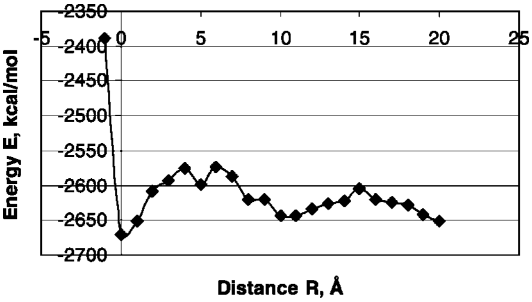


Figure 6. Energy *E* of XIAP-BIR3 interaction with caspase-9 as a function of the distance *R* between the molecules. Computed with the polarizable force field (PFF).

Therefore, the amplitude of the forced motion of the protein and ligand with respect to their equilibrium positions in the complex is

$$y_0 = \frac{a_0}{\omega_0^2 - \omega^2} = P_0 \frac{A_1/M_1 + A_2/M_2}{k/\mu - (2\pi\nu)^2} \tag{8}$$

where ν is the linear frequency of the ultrasound.

Therefore, all we have to do is to (i) calculate the k and frequency $w_0 = (k/\mu)^{1/2}$ of the protein–ligand complex (using the energy vs distance dependence obtained as outlined above and assuming quadratic behavior near the energy minimum) and (ii) find the amplitude of oscillation for the complex y_0 and (iii) check if the complex would be destroyed–or considerably weakened–with such a deviation from the equilibrium distance. Again, the model is crude, but it permits qualitative understanding of the process of ultrasound interaction with the protein–ligand complexes.

III. Results and Discussion

A. Intermolecular Interaction Energies. Interaction energies of the BIR3 domain of XIAP with the caspase-9 protein and a small molecular antagonist were calculated as a function of the separation distances between the XIAP and caspase-9 or the antagonist, as described in section II above. Let us first consider the XIAP-caspase-9 complex. Figures 5 and 6 show these energy profiles calculated with the OPLS-AA and PFF force fields, respectively.

For these and for all the following graphs the point $R = 0$ corresponds to the intermolecular distance found in the original PDB file, and the simulation points are separated

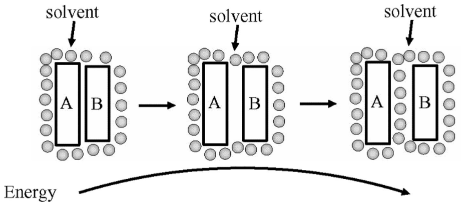


Figure 7. Schematic illustration of the total energy behavior in the process of separating interacting molecules A and B in solution.

by 1.0 Å distance. The lines added to the graphs are simply smoothened lines connecting the points and not a specific interpolation.

Several observations can be made from the above graphs. First of all, both OPLS-AA and PFF predict the intermolecular interaction energy minima at the same point as the experimentally reported structure in the 1NW9 PDB file (within the 1.0 Å precision). Second, the polarizable force field predicts a much steeper growth of the energy for distances shorter than at the equilibrium.

Both Figures 5 and 6 share one important characteristic. As the molecules start getting separated, the energy first increases by several tens of kcal/mol and then drops down, with the total energy of the binding of ca. 30–40 kcal/mol. This effect is well known and is schematically shown in Figure 7. When macromolecules A and B form a complex in solution (left), each of them loses some solvation energy as a part of its surface is not accessible to the solvent. At the same time, some favorable interaction energy is gained because of the intermolecular attractions. When the molecules start separating (center), this dimerization energy becomes less negative, as the distance between the molecules increases. At the same time, the solvent still cannot penetrate the empty space between the molecules A and B, and thus the solvation energy does not become more negative.

Finally, when the molecules are sufficiently far away from each other (right), the solvent can completely solvate both molecules, and the total energy of the complex goes down again, since the lost dimerization energy is at least partially compensated by the energy of solvation. Therefore, the whole process requires transition over an activation barrier. This behavior is precisely what is reproduced by our energy minimizations in continuum solvent, as shown in Figures 5 and 6.

Only a relatively small part of the potential energy curve, with the distances shorter than ca. 4–5 Å, is actually relevant for our efforts to find the harmonic force constant k of the XIAP-caspase complexes as approximated by Figure 4 and consequent equations. Figures 8 and 9 demonstrate the energy-distance dependences for these areas as computed with the OPLS-AA and PFF, respectively.

These two graphs are shown in the same scale, and it can be immediately noticed that the general potential well shape is rather similar with both force fields, even though the approximated curvatures are somewhat different.

In this case, the lines shown on the graphs are results of a polynomial fit. To account for the unharmonicity and to efficiently separate the quadratic form, a fourth degree polynomial was used in each case. The best-fit polynomial for the OPLS-AA force field was $E(R) = 1.1684R^4 - 12.844R^3 + 48.604R^2 - 40.237R - 3761.4$. The PFF polynomial was $E(R) = 6.4187R^4 - 53.7R^3 + 138.1R^2 - 80.635R - 2667.7$. The nonzero linear terms are employed to account for a slight difference between the positions of the experimental ($R = 0$) and calculated energy minima.

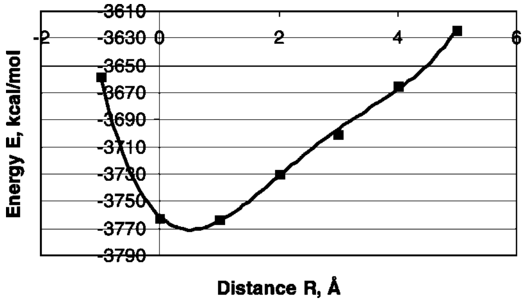


Figure 8. Energy E of the XIAP-BIR3 interaction with caspase-9 as a function of the distance R between the molecules (for small distances). Computed with the OPLS-AA force field.

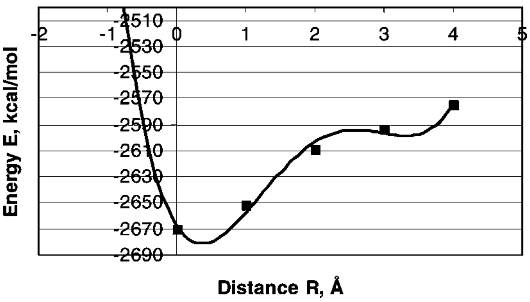


Figure 9. Energy E of the XIAP-BIR3 interaction with caspase-9 as a function of the distance R between the molecules (for small distances) with the polarizable force field (PFF).

The force constants k, as computed from the above equations and the general harmonic formula $E(R) = ½\, kR^2$, are 97.21 kcal/(mol•Å^2) and 276.2 kcal/(mol•Å^2) or 67.58 N/m and 192.0 N/m as computed with the OPLS-AA and PFF, respectively. These values will be used in the next subsection to estimate the effect of an ultrasound irradiation upon the complexes.

Let us now consider the second complex–the one between the BIR3 domain of XIAP and an antagonist to the XIAP-caspase interaction.[23] First of all, it should be pointed out that a successful antagonist has to have a strong interaction energy with the XIAP molecule, and this trend should be reflected in the computational results. Figures 10 and 11 show the dependence of the XIAP-antagonist binding energy on the distance between the two molecules. The line connecting the data points is a smooth connecting line, and $R = 0$ corresponds to the experimentally observed PDB structure (1TFQ).

The following observations can be made here. First, both OPLS-AA and PFF predict the global energy minimum to be at a distance 1.5–2.0 Å shorter than in the PDB structure. Second, the PFF energy is growing steeper than the OPLS-AA one as the molecules get closer to each other–just like

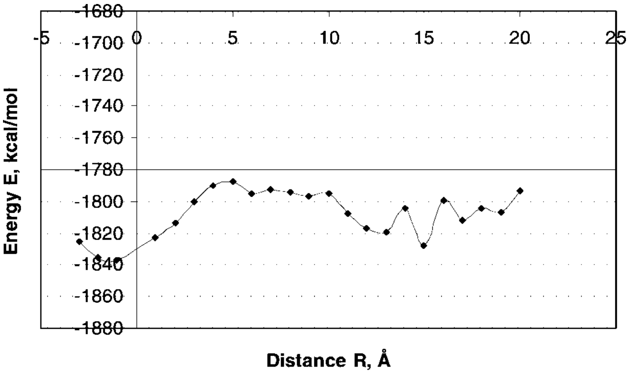


Figure 10. Energy E of the XIAP-BIR3 interaction with the antagonist (shown in Figure 1)[23] as a function of the distance R between the molecules. Computed with the OPLS-AA force field.

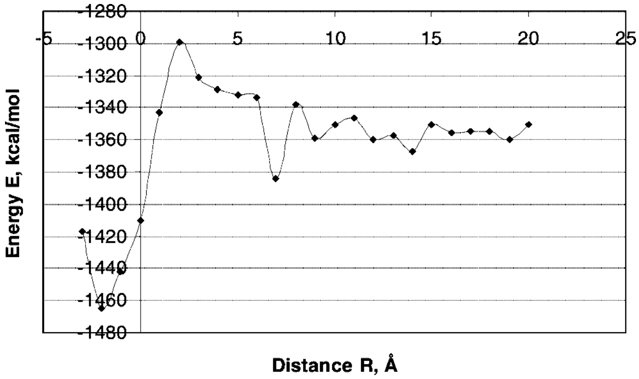


Figure 11. Energy E of the XIAP-BIR3 interaction with the antagonist (shown in Figure 1) as a function of the distance R between the molecules. Computed with the polarizable force field (PFF).

in the case of the XIAP-caspase complex. Third, the general trend of the energy going up and them down as the molecules are separated is still preserved in this case. However, there is a major difference between the PFF and OPLS-AA performance in this case. The binding energy for the complex is about 100 kcal/mol with the former and ca. 40 kcal/mol with the latter. Therefore, PFF predicts the antagonist to be more successful than the OPLS, which only indicates a binding energy as roughly the same as for the XIAP-caspase complex. This pronounced difference can be attributed to the ability of the polarizable force field to react adequately to changes in the electrostatic environment and to interfaces between areas with different dielectric constants (water–protein–ligand). This ligand is smaller than the caspase-9 protein, and thus a sharper adjustment is required, with the molecules experiencing influences of all the three areas. This is why the advantage of the PFF model (more favorable binding energy) seems to be more noticeable in this case of a relatively small molecule bound to the XIAP protein.

Let us now consider the energy-distance dependence for the small distances and derive the strength constants k for the complex. The energy dependence on the intermolecular distances is shown in Figures 12 and 13 for the OPLS-AA and PFF, respectively.

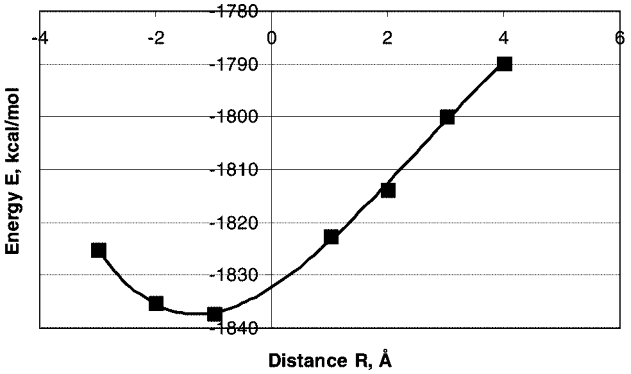


Figure 12. Energy E of the XIAP-BIR3 interaction with the antagonist (shown in Figure 1)[23] as a function of the distance R between the molecules (for small distances). Computed with the OPLS-AA force field.

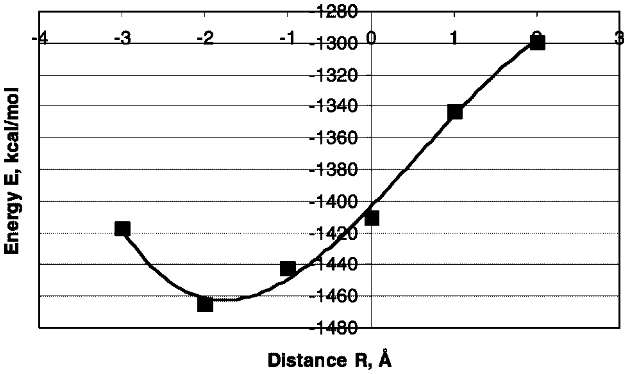


Figure 13. Energy E of XIAP-BIR3 interaction with the antagonist (shown in Figure 1)[23] as a function of the distance R between the molecules (for small distances). Computed with the polarizable force field (PFF).

These graphs are given in different energy scales, as the overall energy changes are greater with the polarizable force field. The lines on the graphs represent fourth degree polynomial fits produced for finding the quadratic part of the energy-distance relationship. The best-fit fourth degree equations were produced in the same way as for the XIAP-caspase complexes. The equations for the OPLS-AA and PFF curves are, respectively, $E(R) = 0.0128R^4 – 0.3361R^3 + 2.0327R^2 + 7.0849R - 1832.3$ and $E(R) = 0.0235R^4 – 3.5619R^3 + 5.858R^2 + 55.013R - 1403.6$, where E is in kcal/mol and R is in Angstroms. This leads to the strength constant k values computed with the OPLS-AA and PFF being 4.066 kcal/(mol·Å^2) and 11.72 kcal/(mol·Å^2), respectively. These values translate into 2.827 N/m and 8.145 N/m. As could be expected, the PFF strength constant is higher, and the difference is greater than for the XIAP-caspase case, for which the OPLS-AA and PFF results are more similar.

The values of the computed k constants are shown together in Table 1.

B. Ultrasound Irradiation Effect on the Strength of the Complexes of XIAP-BIR3. As all the required values of k have been obtained as described above. Let us now list the other parameters required in eq 8. Once all the values are known, this equation can be used to determine the amplitudes of the molecular motion in the protein–ligand complexes and thus estimate their stability.

Table 1. Values of the Strength Constant *k* for the Protein−Ligand Complexes Computed with the OPLS-AA and PFF Force Fields

	XIAP-BIR3 with caspase-9		XIAP-BIR3 with the small antagonist	
complex/method	kcal/(mol·Å^2)	*N*/m	kcal/(mol·Å^2)	*N*/m
OPLS-AA	97.21	67.58	4.066	2.827
PFF	276.2	192.0	11.72	8.145

Table 2. Masses of the Molecules (in amu), Diameters (in Å), and Area ($A = \pi d/4$, in A^2)

value/molecule	mass	diameter	area
caspase-9	30434.13	64.1	3227.05
XIAP-BIR3	13474.28	66.4	3462.79
small antagonist	442.59	16.5	213.82

The masses and the diameters (estimated as the largest atom−atom distance within the molecule) for the caspase-9, XIAP-BIR3, and the antagonist from the PDB 1TFQ are given in Table 2.

The following typical values of the ultrasound frequency and pressure amplitude are assumed: 0.68 MHz and 1.5 MPa.

Having obtained and chosen all these input data, let us now apply eq 8 to find the maximum displacements of the molecules from their equilibrium positions. For the XIAP-caspase complex, the values of y_0 for the OPLS-AA and PFF are, respectively, y_0(OPLS) = 7.525 × 10^{-13} m = 0.007525 Å and y_0(PFF) = 2.649 × 10^{-13} m = 0.002649 Å. For the XIAP-antagonist complex, eq 8 gives the amplitudes of displacements of the molecules with respect to each other of y_0(OPLS) = 1.683 × 10^{-12} m = 0.01683 Å and y_0(PFF) = 5.841 × 10^{-13} m = 0.005841 Å.

It is quite obvious that such small displacements will not be able to change the stability of the complexes in any significant degree. The reason for this result is in the huge differences between the frequency of the ultrasound and the vibrational frequencies of the complexes. These frequencies, entering as $\omega^2 = k/\mu$ into eq 8, are as follows. For the XIAP-caspase complex, the OPLS-AA frequency is $\omega = 2.088 \times 10^{12}$ rad/s or 3.222×10^{11} Hz. The PFF results for this complex are $\omega = 3.519 \times 10^{12}$ rad/s or 5.600×10^{11} Hz. For the XIAP-antagonist complex, the OPLS-AA frequency is $\omega = 1.993 \times 10^{12}$ rad/s or 3.172×10^{11} Hz, and the PFF frequency is $\omega = 3.383 \times 10^{12}$ rad/s or 5.384×10^{11} Hz. Clearly, such a huge difference in the frequencies does not allow the complexes to be anywhere near the resonance while irradiated with the ultrasound.

The above results permit a clear and definite conclusion that the simple frequency resonance hypothesis cannot explain the removal of the apoptosis inhibition by irradiating a tissue with low-intensity ultrasound. It is known however, that a difference in the frequencies can be compensated by using a larger magnitude of the driving oscillations. In this case, we are talking about oscillations of the pressure. It has been shown, both experimentally and theoretically, that irradiation of a liquid with low-intensity ultrasound leads to cavitation, and the bubbles present as a result of cavitation collapse releasing shock waves with pressures of up to 40–60 kbar.[12,13] This leads to the actual pressure amplitude being increased from the nominal 1.5 MPa to ca. 50 kbar = 5 × 10^9 Pa = 5000 MPa. The variation of pressure is no longer obeying the exact cos(ωt) form, but let us assume that eq 8 is still valid for the purpose of estimating the effect. In this case, the amplitude of motion of the caspase-9 and XIAP-BIR3 molecules with respect to each other are 25.1 Å with the OPLS-AA and 8.83 Å with the polarizable force field. The displacement of the XIAP-BIR3 and small antagonist molecules are 56.1 Å with the OPLS-AA and 19.5 Å. Therefore, all the complexes are effectively destroyed if we consider the cavitation resulting from the ultrasound irradiation. We can thus assume that the effect of the ultrasound on the apoptosis reactivation is much more likely to be rooted in the accompanying cavitation and definitely not in the simple frequency resonance. All the calculated displacements and frequencies are shown together in Table 3.

This first of the above results is not a big surprise. A typical ultrasound frequency is much lower than a typical molecular-scale oscillations frequency. Therefore, one could guess that resonance is not responsible for destruction of these molecular complexes without carrying out the detailed molecular calculations described above. However, these calculations are needed to assess the displacement resulting from the ultrasound irradiation quantitatively. A difference in frequencies in driven oscillations can be compensated by a difference in amplitudes–the typical amplitude of ultrasound oscillations is by far greater than the few angstroms of separation needed to destroy a molecular complex. Whether the complex will be destroyed or not is determined by the balance of the unfavorable difference in the frequencies and favorable difference in the amplitudes. And the detailed calculations are needed to quantitatively study this balance. Therefore, our conclusion that, in the presence of cavitation, the ultrasound irradiation is sufficient for destruction of the XIAP complexes, is not trivial and could not be made a priori without the quantitative assessment.

Table 3. Frequences of the One-Dimensional Complex Vibrations (ω in rad/s, N in Hz) and Amplitudes of Changes in the Intermolecular Distances, Å

	frequency		amplitude of displacement	
value/complex	ω	*v*	w/o cavitation	with cavitation
XIAP-caspase,OPLS	2.088×10^{12}	3.322×10^{11}	0.007525	25.1
XIAP-caspase, PFF	3.519×10^{12}	5.600×10^{11}	0.002649	8.83
XIAP-antagonist, OPLS	1.993×10^{12}	3.172×10^{11}	0.01683	56.1
XIAP-antagonist, PFF	3.383×10^{12}	5.384×10^{11}	0.005841	19.5

IV. Conclusions

Interactions of XIAP-BIR3 with caspase-9 and a small antagonist have been studied with the fixed-charges OPLS-AA and polarizable force field (PFF). Energies of the complexes have been calculated as a function of distance. Effects of low-ultrasound irradiation on the strength of the complexes have been assessed with a mechanistic model. It has been found that the polarizable force field predicts steeper walls of the potential wells for the formation of the complexes. In the case of the caspase-9 complex with XIAP, the total energy of the complex formation is ca. 30–40 kcal/mol, as predicted by the both force fields. For the XIAP-antagonist complex, PFF predicts a more negative energy of complex formation, which is consistent with the experimental findings. Both OPLS-AA and PFF reproduce well the increase of the total energy followed by the energy drop, as the molecules are separated from each other in aqueous solution. In general, the results demonstrate that the polarizable force field employed not only is adequate in simulating protein–ligand complexes in solutions but also gives a prediction of a stronger success of the antagonist to the caspase-XIAP interactions.

Estimation of the effect of low-intensity ultrasound on the strength of the complexes demonstrates that the simple frequency resonance hypothesis for the ultrasound-induced reactivation of apoptosis is ruled out. However, the pressure created by the cavitation accompanying the ultrasound irradiation is found to be sufficient to destroy the caspase-9 inhibition and, as a result, is named as the most probable candidate for the mechanism of apoptosis reactivation. While the overall mechanistic model of the ultrasound-molecular complex interaction is crude, it permits a qualitative explanation of the experimentally observed phenomena.

Acknowledgment. This work was supported in part by the Central Michigan University Faculty Research and Creative Endeavors (FRCE) grant 48984.

References

(1) (a) Shi, Y. *Protein Sci*. 2004, *13*, 1979. (b) Shiozaki, E. N.; Chai, J.; Rigotti, D. J.; Riedl, S. J.; Alnemri, E. S.; Fairman, R.; Shi, Y. *Mol. Cell* **2003**, *11*, 519.

(2) (a) See, for example: (a) Igney, F. H.; Krammer, P. H. *Nat. Rev. Cancer* **2002**, *2*, 277. (b) Reed, J. C. *Nat. Rev. Drug Discovery* **2002**, *1*, 111. (c) Los, M.; Burek, C. J.; Stroh, C.; Benedyk, K.; Hug, H.; Mackiewicz, A. *Drug Discovery Today* **2003**, *8*, 67.

(3) (a) Takahashi, R.; Deveraux, Q.; Tamm, I.; Welsh, K.; Assa-Munt, N.; Salvesen, G. S.; Reed, J. C. *J. Biol. Chem.* **1998**, *273*, 7787. (b) Riedl, S. J.; Renatus, M.; Schwarzenbacher, R.; Zhou, Q. Sun, C., Fesik, S. W.; Liddington, R. C.; Salvesen, G. S. *Cell* **2001**, *104*, 791. (c) Holcik, M.; Gibson, H.; Korneluk, R. G. *Apoptosis* **2001**, *6*, 253–261.

(4) (a) Liu, Z.; Sun, C.; Olejniczak, E. T.; Meadows, R. P.; Betz, S. F.; Oost, T.; Herrmann, J.; Wu, J. C.; Fesik, S. W. *Nature* **2000**, *408*, 1004. (b) Shi, Y. *Cell Death Differ.* **2002**, *9*, 93. (c) Du, C.; Fang, M.; Li, Y.; Li, L.; Wang, X. *Cell* **2000**, *102*, 33. (d) Verhagen, A. M.; Ekert, P. G.; Pakusch, M.; Silke, J.; Vaux, D. L. *Cell* **2000**, *102*, 43. (e) Nikolovska-Coleska, Z.; Xu, L.; Hu, Z.; Tomita, Y.; Li, P.; Roller, P. P.; Wang, R.; Fang, X.; Guo, R.; Zhang, M.; Lippman, M. E.; Yang, D.; Wang, S. *J. Med. Chem.* **2004**, *47*, 2430. (f) Schimmer, A. D. *Cancer Res.* **2004**, *64*, 7183.

(5) Johns, L. D. *J. Athletic Training* **2002**, *37*, 291.

(6) Mitragotri, S. *Nat. Rev. Drug Discovery* **2005**, *4*, 255.

(7) Yu, T.; Wang, Z.; Mason, T. J. *Ultrason. Sonochem.* **2004**, *11*, 95.

(8) Kondo, T.; Feril, L. B. *J. Radiat. Res.* **2004**, *45*, 479.

(9) Mohamed, M. M.; Mohamed, M. A.; Fikry, N. M. *Ultrasound Med. Biol.* **2003**, *29*, 1635.

(10) Lagneaux, L.; Meulenaer, E. C.; Delforge, A.; Dejeneffe, M.; Massy, M.; Moerman, C.; Hannecart, B.; Canivet, Y.; Lepeltier, M.-F.; Bron, D. *Exp. Hematol.* **2002**, *30*, 1293.

(11) Yu, T.; Xiong, Z.; Chen, S.; Tu, G. *Ultrason. Somochem.* **2005**, *12*, 345.

(12) Wu, C. C.; Roberts, P. H. *Phys. Rev. Lett.* **1993**, *70*, 3424.

(13) Pecha, R.; Gompf, B. *Phys. Rev. Lett.* **2000**, *84*, 1328.

(14) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(15) Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 730. and references therein.

(16) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(17) Dang, L. X.; Chang, T.-M. *J. Chem. Phys.* **1997**, *106*, 8149.

(18) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.

(19) For representative publications, see:(a) Liu, Y. P.; Kim, K.; Berne, B. J.; Friesner, R. A.; Rick, S. W. *J. Chem. Phys.* **1998**, *108*, 4739. (b) Yu, H. B.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221. (c) Jardon-Valadez, E.; Costas, M. E. *THEOCHEM* **2004**, *677*, 227. (d) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1. (e) Maple, J. R.; Cao, Y. X.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694.

(20) Kaminski, G. A. *J. Phys. Chem. B* **2005**, *109*, 5884.

(21) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.

(22) *Maestro version 5.0*; Schrödinger, Inc.: Portland, OR, 2002.

(23) Oost, T. K.; Sun, C.; Armstrong, R. C.; Al-Assaad, A.-S.; Betz, S. F.; Deckwerth, T. L.; Ding, H.; Elmore, S. W.; Meadows, R. P.; Olejniczak, E. T.; Oleksijew, A.; Oltersdorf, T.; Rosenberg, S. H.; Shoemaker, A. R.; Tomaselli, K. J.; Zou, H.; Fesik, S. W. *J. Med. Chem.* **2004**, *47*, 4417.

(24) *Impact version 3.6*; Schrödinger, LLC: Portland, OR, 2005.

(25) (a) See, for example: (a) Naruse, Y. *Biosystems* **2002**, *66*, 55. (b) Naruse, Y. *Jpn. J. Appl. Phys.* **2004**, *43*, 3629.

CT8000188

JCTC Journal of Chemical Theory and Computation

Prediction of Protein Loop Conformations Using the AGBNP Implicit Solvent Model and Torsion Angle Sampling

Anthony K. Felts,[†] Emilio Gallicchio,[†] Dmitriy Chekmarev,[†] Kristina A. Paris,[†]
Richard A. Friesner,[‡] and Ronald M. Levy*[,†]

Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey 08854, and Department of Chemistry, Columbia University, New York, New York 10027

Received February 19, 2008

Abstract: The OPLS-AA all-atom force field and the Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model, in conjunction with torsion angle conformational search protocols based on the Protein Local Optimization Program (PLOP), are shown to be effective in predicting the native conformations of 57 9-residue and 35 13-residue loops of a diverse series of proteins with low sequence identity. The novel nonpolar solvation free energy estimator implemented in AGBNP augmented by correction terms aimed at reducing the occurrence of ion pairing are important to achieve the best prediction accuracy. Extended versions of the previously developed PLOP-based conformational search schemes based on calculations in the crystal environment are reported that are suitable for application to loop homology modeling without the crystal environment. Our results suggest that in general the loop backbone conformation is not strongly influenced by crystal packing. The application of the temperature Replica Exchange Molecular Dynamics (T-REMD) sampling method for a few examples where PLOP sampling is insufficient are also reported. The results reported indicate that the OPLS-AA/AGBNP effective potential is suitable for high-resolution modeling of proteins in the final stages of homology modeling and/or protein crystallographic refinement.

1. Introduction

A necessary component for an effective computational approach to the homology modeling problem[1] for protein structure prediction[2] and crystallographic and NMR structure refinement[3,4] is a scoring function that scores more favorably the native conformation over other possible conformations.[5,6] Scoring functions aimed at fold recognition and secondary structure assignment have been evaluated on the basis of their ability to recognize the known native protein conformation among a set of plausible misfolded decoy structures.[7–12] Both physics-based[13–19] and empirical knowledge-based scoring functions[20–23] have performed reasonably well in this kind of evaluation tests.

Recent development efforts have been focused on the refinement stages of the homology modeling problem, such as the conformational prediction of protein loops[24–26] and surface side chains[27] as well as the modeling of ligand/receptor induced fit effects,[28] which are essential steps to make the model useful as a drug discovery and optimization target. These kinds of high-resolution protein structure prediction applications have generally been performed using atomistic physics-based free energy estimators.

Protein decoy scoring exercises have been useful in determining the key global features of physics-based energy functions (such as the inclusion of solvation effects)[19] necessary for recognizing the broad characteristics of native protein structures. The decoy evaluation technique, however, is in general too blunt an instrument for discriminating the ability of energy functions to recognize small structural variations within the native ensemble. For thorough testing,

* Corresponding author e-mail: ronlevy@lutece.rutgers.edu.
[†] Rutgers University.
[‡] Columbia University.

10.1021/ct800051k CCC: $40.75 © 2008 American Chemical Society
Published on Web 04/26/2008

it is necessary to challenge the energy function by performing extensive local conformational searches to actively look for minima of the energy functions and measure the degree of correspondence of these with the known native conformation.

Determining the correct conformation of a loop on a protein is one of the final steps in homology model building. After secondary structures have been assigned and placed, model construction often proceeds by conformational prediction of connecting loops. In loop prediction tests, we assume that the rest of the protein frame has been folded accurately and the conformation of the loop of interest remains to be determined. Effectively, the loop is a tethered peptide whose conformations can be sampled extensively while in the presence of the energy field generated by the rest of the protein. Many different conformations of the loop can be generated and tested for false global minima which exist in the presence of the effective potential field of the protein framework. This makes the protein loop prediction problem a powerful benchmarking tool to test the accuracy of energy functions.

An accurate molecular mechanics model suitable for protein structure prediction and refinement requires a representation of the aqueous solvent environment. The polarization of the solvent favors the hydration of polar and especially charged groups that, in the absence of solvation forces, tend to form non-native intramolecular interactions. Explicit solvent models provide the most detailed and complete description of hydration phenomena.[29] However, computer simulations using explicit solvent models are computationally intensive, not only just because of the much larger number of atomic interactions that need to be considered but also, and perhaps more importantly, because of the need to average the fluctuating effects of the solvent reaction field to obtain a meaningful estimate of the solvation free energy of each protein conformation. For protein structure prediction applications effective potential models that treat the solvent implicitly have much to offer. The modeling community has developed a strong interest in a class of implicit solvent models based on the Generalized Born framework;[30–32] an approximation of the Poisson equation of continuum electrostatics.[33,34] Much of the popularity of Generalized Born (GB) models stems from their computational efficiency and ease of integration in molecular simulation computer programs.[31,35–38] Generalized Born models have been shown to be able to reproduce with good accuracy Poisson[32,39–41] and explicit solvent[42,43] results at a fraction of the computational expense.

In this work we evaluate the accuracy of the Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model,[44] in predicting the native conformation of protein loops using the Protein Local Optimization Program (PLOP).[26] The PLOP program[26] performs loop and side chain conformational predictions based on an efficient hierarchical conformational sampling algorithm in torsional angle space, combined with a recent parametrization of the OPLS-AA force field[45,46] and a Generalized Born implicit solvation model. The AGBNP implicit solvent model is based on an analytical pairwise descreening[47] implementation of the Generalized Born model[30] and a novel nonpolar hydration free energy model which combines separate estimators for the solute−solvent van der Waals dispersion energy and the work of cavity formation.[48–50]

We previously showed[44] that the OPLS-AA/AGBNP effective potential was able to consistently score native loop conformations more favorably than non-native decoy loop conformations generated by PLOP using the OPLS-AA/SGB/NP effective potential.[26] The present work extends that work by including a larger set of loops as well as longer loops targets and by employing the OPLS-AA/AGBNP model directly in the conformational search and optimization procedure implemented in PLOP. We also evaluate various parametrizations of the AGBNP model to determine the role of the nonpolar model and of the correction terms we developed aimed at reducing the occurrence of intramolecular ion pairing, and we compare them to the distance dependent dielectric and the Surface Generalized Born (SGB/NP)[51,52] solvation models as implemented in the PLOP program.

As part of this work we have also evaluated the efficiency of the recently proposed loop conformational search schemes based on PLOP[26,53] which improves on earlier torsion angle based sampling methods.[24,25] These PLOP-based conformational search schemes have been optimized for loop conformational prediction in the crystal environment. We evaluate enhanced versions of these schemes more suitable for loop prediction calculations in the solution environment (the biologically relevant environment for most homology modeling applications). We also tested the applicability of temperature Replica Exchange Molecular Dynamics (T-REMD) to the problem of protein loop prediction, which, given its favorable scaling with respect to the number of degrees of freedom, offers an alternative route for conformational prediction of long loops and for simultaneous refinement of interacting protein elements.

2. Methods

2.1. Loop Prediction Algorithms. The loop prediction algorithm implemented in the Protein Local Optimization Program (PLOP) is described in detail in ref 26. During loop buildup, a series of filters of increasing complexity is applied to eliminate unreasonable conformations as early as possible. Some of these filters detect clashes between backbone atoms and the atoms of the rest of the protein (referred to as the frame) and check that enough space is available to place the side chain of each residue. On the order of hundreds to thousands of loop conformations are generated in the loop build-up stage. To reduce the number of conformations passed to the next stages, loop conformations are clustered based on backbone rmsd using the K-means algorithm,[54] a clustering method that requires a predetermined number of clusters. The two most important parameters that control the tradeoff between accuracy and efficiency of PLOP's loop prediction algorithm are the overlap factor parameter (*ofac*), defined as the minimum permitted ratio of the interatomic distance over the sum of the Lennard-Jones radii of the atoms of interest, which controls the amount of overlap tolerated between any two atoms, and the number of clusters N_{clust}. A smaller *ofac* allows more overlap between atoms which in effect allows for more loop conformations to be sampled which otherwise would have been eliminated due to steric clashes. The efficiency of the loop-prediction procedure is partially determined by the value of *ofac*. If *ofac* is too small, a large number of irrelavent loop conformations are generated that have to be processed in subsequent steps. On the other hand with a large *ofac* nativelike loops may be rejected due to steric clashes caused by the discreteness of the torsion library used to generate the loops. Based on the oberved value of *ofac* found in the PDB, Jacobson et al. set *ofac* to between 0.70 and 0.75.[26] The number of clusters N_{clust} needs to be sufficiently large to account for each nonredundant loop conformation. If N_{clust} is set too small, conformationally different structures could potentially be clustered together. The number of nonredundant loop conformations will depend upon how large is the conformational space available to the loop. Based on empirical evidence, Jacobson et al. set the number of clusters to four times the number of residues in the loop.[26]

The PLOP program allows sampling of loop conformations in the crystalline phase with the SGB/NP solvation model.[42,42] We performed SGB/NP prediction calculations with and without crystal symmetry in order to compare with previous literature.[26] Loop prediction calculations with all of the other implicit solvent models were conducted without crystal symmetry.

The basic loop prediction algorithm described above is often insufficient for loops with nine or more residues. For these longer loops we have adopted prediction schemes based on multiple executions of PLOP with different parameters.[26,53] These schemes are based on focusing conformational sampling in promising and progressively smaller regions of conformational space. The initial predictions with the most favorable energy scores are subjected to a series of constrained refinement calculations with PLOP in which selected loop backbone atoms are not allowed to move or move only within a given range.

The standard 9-residue loop prediction scheme is based on the procedure described in detail by Jacobson et al.[26] For loops which the standard version of loop prediction fails to find low-energy, nativelike conformations, we attempted to predict these loops with an extended version of the loop prediction algorithm. An extended version of this scheme involves using twice the number of clusters (from 36 to 72) and reduced *ofac* (overlap factor) coefficients (0.5 instead of 0.75) during the initial prediction stage. All other stages as described by Jacobson et al.[26] remain the same.

For the 13-residue loops we have adopted an alternative long loops prediction scheme developed previously for longer loops.[53] This scheme is based on the idea of refining the loop structure by sampling increasingly shorter loop segments which can be handled by PLOP's conformational search procedure. Briefly, initial predictions are produced with 3 different overlap factors (0.65, 0.70, and 0.75) and subjected to constrained refinement. The five lowest-energy nonredundant structures so obtained are passed to a series of loop prediction stages which sample progressively shorter segments obtained by fixing any possible combination up to five residues at either terminal end of the loop. The *standard sampling* and *extended sampling* variations of this sampling method differ in the number of nonredundant lowest-energy models that are processed at each stage. With extended sampling five lowest-energy models are passed from one stage to the next. With standard sampling the number of PLOP iterations is reduced by half by progressively reducing the number of models passed to later stages.

Table 1. 9-Residue and 13-Residue Loops Indicated by Their the Protein Data Bank (PDB) Designation for the Protein and R_{first} and R_{last} Are, Respectively, The First and Last Residue of the Loop[a]

PDB(R_{first} - R_{last})	PDB(R_{first} - R_{last})	PDB(R_{first} - R_{last})
1aac(58−66)	1pda(108−116)	1cnv(110−122)
1aba(69−77)	1pgs(117−125)	1d0c(A:280−292)
1amp(57−65)	1php(91−99)	1dpg(A:352−364)
1arb(90−98)	1ptf(10−18)	1dys(A:290−302)
1arb(168−176)	1ra9(142−150)	1ed8(A:67−79)
1arp(127−135)	1rhs(216−224)	1eok(A:147−159)
1aru(36−44)	1sgp(E109−E117)	1f46(A:64−76)
1btl(102−110)	1tca(170−178)	1g8f(A:72−84)
1byb(246−254)	1tca(217−225)	1gpi(A:308−320)
1cse(E95−E103)	1xif(59−67)	1h4a(X:19−31)
1csh(252−260)	1xnb(116−124)	1hnj(A:191−203)
1ede(257−265)	1xyz(A568−A576)	1hxh(A:87−99)
1fus(31−39)	1xyz(A795−A803)	1iir(A:197−209)
1fus(91−99)	1xnb(133−141)	1jp4(A:153−165)
1gpr(63−71)	1wer(942−950)	1kbl(A:793−805)
1isu(A30−A38)	2alp(139−159)	1krh(A:131−143)
1ivd(244−252)	2ayh(169−177)	1l8a(A:691−703)
1lkk(A142−A150)	2cpl(24−32)	1lki(62−74)
1lkk(A193−A201)	2eng(172−180)	1m3s(A:68−80)
1mla(194−202)	2hbg(18−26)	1mo9(A:107−119)
1mrj(92−100)	2sil(183−191)	1nln(A:26−38)
1mrk(53−61)	3pte(78−86)	1o6l(A:386−398)
1mrp(284−292)	3pte(107−115)	1ock(A:43−55)
1nfp(12−20)	3pte(215−223)	1ojq(A:167−179)
1nif(266−274)	3tgl(56−64)	1p1m(A:327−339)
1nls(131−139)	4gcr(94−102)	1qqp(2:161−173)
1noa(9−17)	1a8d(155−167)	1qs1(A:389−401)
1noa(76−84)	1ako(203−215)	1xyz(A:645−657)
1noa(99−107)	1arb(182−194)	2hlc(A:91−103)
1npk(102−110)	1bhe(121−133)	2ptd(136−148)
1onc(70−78)	1bkp(A:51−63)	

[a] A letter indicates the chain on which the loop is found.

We also investigated if a technique based on replica exchange molecular dynamics importance sampling could predict loop conformations. We selected 9-residue loops which were not successfully predicted by the standard sampling algorithm built around PLOP to see if importance sampling would succeed. This subset of the 9-residue loops (Table 3) was investigated with the temperature replica exchange sampling method (T-REMD)[55–57] as implemented in the IMPACT software package.[57] The lowest-energy loop configuration obtained in the third stage of PLOP optimization was chosen as a starting point for the corresponding T-REMD run. Each loop was minimized in the field of the surrounding immobilized protein frame. T-REMD was based on constant temperature MD, and exchanges between replicas were attempted every 500 steps. During T-REMD simulations, the protein frame conformation was fixed. The OPLS-AA force field was employed to model the intramolecular potential, while the solvent was treated implicitly by the AGBNP+ effective potential model (see below). We used 12 replicas at 270, 298, 329, 363, 401, 442, 488, 539, 595,

Table 2. Summary of the Loop Conformational Predictions Results with the Standard and Enhanced Sampling Procedures[a]

	9-residue						
	SGB/NP-X	SGB/NP	ddd	AGB-γ	AGBNP	AGBNP+	13-residue AGBNP+
E	8	11(10)	19	6	4(3)	2	2
S	5(5)	7(14)	4(7)	4(7)	4(9)	5(10)	5(14)
M	2	3(4)	3	1	0	1	1(2)
E+*S*+*M*	15	21	26	11	8	8	8
⟨rmsd⟩	1.44	1.91	2.31	1.10	1.04	1.00	1.87
median rmsd	0.58	0.60	1.27	0.52	0.52	0.58	0.67

[a] SGB/NP-X: SGB/NP with crystal symmetry; ddd: distance-dependent dielectric; *E*: number of energy errors (results listed for both enhanced and (standard) sampling); *S*: number of sampling errors (results listed for both enhanced and (standard) sampling); *M*: number of marginal errors (results listed for both enhanced and (standard) sampling). The values listed were obtained with enhanced sampling; the values in parentheses were obtained with standard sampling. ⟨rmsd⟩: average rmsd (in Å) of the lowest-energy loops.

Table 3. Summary of the Loop Conformational Predictions Results with the OPLS-AA/AGBNP+ Force Field and T-REMD Conformational Sampling, Compared to the Corresponding Predictions with the PLOP-Based Standard Sampling Procedure

PDB(R_{first} - R_{last})	PLOP rmsd (Å)	T-REMD rmsd (Å)
1npk(102−110)	3.60	4.30
1onc(70−78)	7.43	2.06
1fus(31−39)	6.03	1.78
1byb(246−254)	4.00	4.95
1noa(99−107)	5.67	3.94
1wer(942−950)	4.29	1.34

657, 725, and 800 K. The T-REMD simulation length varied from 15 to 35 ns, and the data collected over the last 5 ns of the T-REMD trajectories were used for final analysis.

2.2. The Energy Functions. The energy functions we used to score the predicted loops are composed of the all-atom force field, OPLS-AA,[45,46] and an implicit solvent model. The particular version of OPLS-AA[46] we used has improved torsional parameters based on fits to high-level LMP2 quantum chemical calculations of the torsion interactions of small peptides. These fits led to improvements in the accuracy of the φ, ψ, and side chain χ torsion energies for amino acids.[27]

The implicit solvent models we investigated in this study are the simple distance-dependent dielectric and two generalized Born solvation models, the Surface Generalized Born (SGB) [42,42] and Analytical Generalized Born (AGB).[44] It is assumed in the distance-dependent dielectric model that the interaction energy between partial charges in a heterogeneous dielectric environment follows a simple Coulomb law. The Coulomb energy term is given by

$$E_{\text{Coul}} = \frac{q_i q_j}{\varepsilon r_{ij}} \tag{1}$$

where r_{ij} is the interatomic distance between atoms i and j, and ε is the dielectric constant. In the distance-dependent dielectric model, ε is no longer constant but proportional to the interatomic distance as such

$$\varepsilon = r_{ij} \tag{2}$$

While the distance-dependent dielectric is known to be a poor model for solvation, we use the results generated with it to benchmark the improvements in loop prediction that can be obtained with more accurate physical models.

2.2.1. SGB/NP Implicit Solvent Model. The SGB model is the surface implementation[42,51] of the generalized Born model.[30] The generalized Born equation

$$G_{\text{GB}} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{w}}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}(r_{ij})} \tag{3}$$

where q_i is the charge of atom i and r_{ij} is the distance between atoms i and j, gives the electrostatic component of the free energy of transfer of a molecule with interior dielectric ε_{in} from vacuum to a continuum medium of dielectric constant ε_{w}, by interpolating between the two extreme cases that can be solved analytically: the one in which the atoms are infinitely separated and the other in which the atoms are completely overlapped. The interpolation function f_{ij} in eq 3 is defined as

$$f_{ij} = \left[r_{ij}^2 + B_i B_j \exp\left(-r_{ij}^2/4B_i B_j\right)\right]^{\frac{1}{2}} \tag{4}$$

where B_i is the Born radius of atom i defined as the effective radius that reproduces through the Born equation

$$G_{\text{single}}^{i} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{w}}}\right)\frac{q_i^2}{B_i} \tag{5}$$

the electrostatic free energy of the molecule when only the charge of atom i is present in the molecular cavity. The G_{single}^{i} are evaluated numerically by integrating the interaction between atom i and the charge induced on the solute−solvent boundary surface, S, by the Coulomb field of this atom

$$G_{\text{single}}^{i} = -\frac{1}{8\pi}\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{w}}}\right)\int_S \frac{q_i^2}{|\mathbf{r} - \mathbf{r}_i|^4}(\mathbf{r} - \mathbf{r}_i)\cdot\mathbf{n}(\mathbf{r})d^2\mathbf{r} \tag{6}$$

where $\mathbf{n}(\mathbf{r})$ is the normal to the surface, S, at $\mathbf{r}$. The atomic radii that define the solute−solvent dielectric boundary are set to the van der Waals radii based on the Lennard-Jones σ parameters. The Born radii for eq 4 are calculated using eqs 5 and 6. In this work, we set $\varepsilon_{\text{in}} = 1$ and $\varepsilon_{\text{w}} = 80$. The SGB implementation used in this work includes further correction terms that bring the SGB reaction field energy into closer agreement with exact PB results.[51] Coupled with the SGB model is a function describing the nonpolar interactions between the solute and solvent which is based on two terms: the van der Waals interaction between solute and solvent and the work to form the cavity in the solvent. The full solvation model is referred to as SGB/NP. Exact details of the nonpolar function in SGB/NP can be found in ref 52.

2.2.2. AGBNP Implicit Solvent Model. The analytical generalized Born (AGB) implicit solvent model differs from SGB in the way that the Born radii are calculated. AGB is based on a novel pairwise descreening implementation[44] of the generalized Born model.[58] The combination of AGB with a recently proposed nonpolar hydration free energy estimator described below is referred to as AGBNP.[44] AGB employs a parameter-free and conformation-dependent analytical scheme to obtain the pairwise descreening scaling coefficients used in the computation of the Born radii used in the generalized Born equation, eq 3. The agreement between the AGB Born radii and exact numerical calculations was found to be excellent.[44] The AGBNP nonpolar model consists of an estimator for the solute−solvent van der Waals interaction energy in addition to an analytical surface area component corresponding to the work of cavity formation.[44] Because AGBNP is fully analytical with first derivatives it is well suited for energy minimization as well as MD sampling. A detailed description of the AGBNP model and its implementation is provided in ref 44.

The nonpolar solvation free energy is given by the sum of two terms: the free energy to form the cavity in solvent filled by the solute and the dispersion attraction between solute and solvent.[49,59] The nonpolar free energy is written as[44]

$$\Delta G_{\mathrm{np}} = \sum_i \left(\gamma_i A_i + \Delta G_{\mathrm{vdW}}^{(i)}\right) \tag{7}$$

where the first term is the cavity term, γ_i, is the surface tension proportionality constant for atom i, and A_i is the solvent exposed surface area of atom i. The second term is the dispersion interaction term which is given by[44]

$$\Delta G_{\mathrm{vdW}}^{(i)} = \alpha_i \frac{-16\pi\rho_w \varepsilon_{i,w}\sigma_{i,w}^6}{3(B_i + R_w)^3} \tag{8}$$

where α_i is an adjustable solute−solvent van der Waals dispersion parameter for atom i. The parameter ρ_w is the number density of water at standard conditions (0.033428/Å ^{3}). $\varepsilon_{i,w}$ and $\sigma_{i,w}$ are the pairwise Lennard-Jones (LJ) well-depth and diameter parameters for atom i and the TIP4P water oxygen as given by the OPLS-AA force field.[45,46] ($\varepsilon_{i,w} = \sqrt{\varepsilon_i \varepsilon_{\mathrm{w},}}$ where ε_i is the LJ well-depth for atom i and ε_w is similarly for the TIP4P water oxygen. The ε for water hydrogens is set to zero. $\sigma_{i,w}$ is defined in a similar manner.) R_w is the radius of a water molecule (1.4 Å). By not incorporating the Lennard-Jones parameters into the dispersion parameter, α_i, atoms with different though similar ε_i's and σ_i's are assigned the same α so as to minimize the number of adjustable parameters. B_i is the Born radius of atom i. The Born radius in this equation provides a measure of how buried atom i is in the solute. The deeper the atom is in the solute, the smaller will be its contribution to the total solute−solvent dispersion interaction energy. The functional form of ΔG_{vdW} in both SGB/NP and AGBNP depends upon the Born radius since it is a measure of the degree of burial of the atom. In SGB/NP, the dependence of ΔG_{vdW} on the Born radius was chosen on an ad hoc basis. The form of eq 8 for the solute−solvent van der Waals interaction energy component has been derived on the basis of simple physical arguments.[44]

In this work we use two sets of parametrizations of α and γ to test the full nonpolar function described above relative to a simpler nonpolar function. In past implementations,[19] the total nonpolar solvation free energy is given by a term proportional to the solvent-accessible surface area, or in terms of eq 7, setting all values of α_i to zero

$$\Delta G_{\mathrm{np}} = \sum_i (\gamma_i A_i) \tag{9}$$

where γ_i is set for all atoms to 0.015 kcal/mol/Å^2. This implicit solvent model with the less-detailed nonpolar function is referred to as "AGB-γ". When we use the full nonpolar function including the dispersion term (eq 8) using the parameters set forth in the work of Gallicchio and Levy,[44] the implicit solvent model is referred to as "AGBNP".

A third parametrization aimed at implementing a correction for salt bridge interactions (which are generally overestimated by generalized Born solvent models)[56,60] is also investigated. To correct for the overstabilization of salt bridges by the generalized Born model, we used modified radii and γ_i for carboxylate oxygens. The radius of the carboxylate oxygen is decreased from 1.48 Å, as in the original AGBNP, to 1.30 Å; γ_i of the carboxylate oxygen is set to −0.313 kcal/mol/Å^2. These have the combined effect of increasing the solubility of carboxylate oxygens and decreasing the likelihood of ion pairing between the carboxylate groups on glutamate and aspartate and positively charged groups found on lysine and arginine. We have parametrized this radius and γ_i to experimental data for small molecules and to provide results which matched those generated with explicit solvent (unpublished results). The implicit solvent model that has additional descreening of ion pairing is referred to as "AGBNP+".

2.3. The Protein Loop Data Sets. We have tested the loop prediction algorithms on two sets of protein loops of known structure of nine and 13 residues in length. The first set is composed of the 57 9-residue loops listed in Table 1. This set was originally compiled by Fiser et al.[24] and by Xiang et al.[25] The 35 13-residue loop set is the same as the one investigated by Zhu et al.[53] These loops were culled from the PISCES[61] database. The proteins in these databases have been filtered using the following selection criteria: (i) low sequence identity (60% for Fiser et al.,[24] 20% for Xiang et al.,[25] and <40% for Zhu et al.),[53] (ii) complete X-ray structure available with resolution <2 Å, R < 0.25, and average temperature factor within the loop <35, (iii) 6.5 < pH < 7.5, (iv) overlap factor for any loop atom >0.7, (v) no significant loop secondary structure, (vi) no more than 4 additional loop residues on either side of the selected loop, (vii) distance between any loop atom and any ligand atom >4 Å (6.5 Å for a metal ion).[26,53] While some of the loops contain very small amounts of secondary structure, in general, they are representative of longer loops found in globular proteins. All crystallographic water molecules are removed prior to loop prediction. Hydrogen atoms are added to each structure.[26]

2.4. Characterization of the Predicted Loop Structures. The predicted loop conformation is the one that has the lowest energy among those found by the conformational search procedures described above. The accuracy of the predicted conformations is analyzed by computing their root-mean-square deviation (rmsd) with respect to the corresponding crystallographically determined native structures (the X-ray structure). The native and predicted protein loops are already in a common frame because only the conformation of the loop is varied during loop torsion angle sampling. The rmsd of the backbone atoms (N, C, and C_α) predicted and X-ray conformations are calculated in this common frame. We characterize the accuracy of the predictions based on the average and median backbone rmsd of the predictions and the number of correct predictions. Correct predictions are defined as those that fall within a chosen rmsd threshold value from the X-ray structure.

An incorrect prediction (one with an rmsd larger than the threshold, see below) is further classified as an *energy error* when the prediction has an energy significantly lower than native, and otherwise as a *sampling error*, when the predicted loop has an energy higher than the native. This classification of incorrect predictions is aimed at determining the cause of the failure of the method to produce a nativelike conformation. An energy error is indicative of the failure of the energy function to score the native conformation more favorably than non-native conformations; so that, even if the conformational search method had produced them, near-native conformations would not be recognized as good predictions. A sampling error is indicative of the conformational search procedure failing to sample conformations near the native conformation, even though the energy function scores at least some of them more favorably than non-native conformations.

The classification of correct and incorrect predictions requires the specification of a rmsd threshold value. This choice depends on the level of prediction accuracy required by the application. We report our results based on C_α rmsd thresholds of 1.5 and 2.0 Å for the 9- and 13-residue loop sets, respectively, which have been used before to analyze the accuracy of loop prediction methods.[26,53] In addition, the classification of incorrect predictions requires the specification of an energy gap threshold value. If the difference in energies of the native and predicted conformations (where the predicted is lower in energy than the native) exceeds the energy gap threshold value, the incorrect prediction is classified as an energy error. In this work the results have been reported using an energy gap threshold value of 5 kcal/mol. The choice of this value absorbs the effects due to configurational entropy missing from our free-energy estimator as well as the acceptable level of error in the energy function. We have explored a range of rmsd and energy gap threshold parameters and confirmed that the conclusions drawn in this work are not qualitatively affected by the particular choices made here. The energy of the native conformation used in the computation of the energy gap of the predicted conformation is determined in three ways: (1) a minimization of the loop with the frame, (2) a minimization followed by an optimization of the side chains on the loop, and (3) a confined search within 2 Å rmsd from the X-ray conformation similarly as for the second stage of refinement in the loop prediction procedure. We selected the native energy as the lowest energy determined from any of these. In almost all cases this conformation differs from the X-ray structure by no more than 1 Å C_α rmsd.

A minority of incorrect predictions were not classifiable as either energy errors or sampling errors. These were typically cases that do not qualify as clear energy errors because, even though the energy of the predicted non-native conformation is lower than the native conformation, the magnitude of the energy gap is within the 5 kcal/mol margin and do not qualify as sampling errors because native conformations of reasonable low energy were sampled. In the following we label these cases as *marginal errors*. Marginal errors are effectively incorrect predictions due to subtle and not easily attributable energetic, entropic, and methodological causes.

In order to be able to compare the T-REMD predictions with those obtained from the PLOP-based prediction schemes and with the native structures, we energy-minimized the loop conformations found at the lowest target temperature of 270 K and recomputed the loop backbone rmsds with respect to the reference crystal structure. The conformation with the lowest energy was selected as the predicted conformation. The predicted conformation was then classified in terms of the energy gap and rmsd from the native conformation using the scheme described above.

3. Results

The results of the loop prediction tests are summarized in Table 2 for the standard and extended conformational sampling procedures (see Methods). Extended sampling was conducted on the loops that resulted in a sampling error with standard sampling; Table 2 includes the combined standard and extended sampling results. For the 57 9-residue loops (see Table 1) loop prediction tests were conducted with OPLS-AA and the following implicit solvent models: distance-dependent dielectric, SGB/NP, AGB-γ, AGBNP, and AGBNP+. It has been stated that the results for loop prediction with PLOP was independent of the presence of crystal symmetry.[26] However, we found that crystal symmetry significantly influenced the results with SGB/NP. In order to compare with previous results,[26] we performed loop predictions with SGB/NP both in the presence and absence of crystal symmetry. Loop prediction calculations with all of the other implicit solvation models were conducted only in the absence of crystal symmetry. Loop prediction tests for the 35 13-residue loops (see Table 1) were conducted with AGBNP+. As described in the Methods section we characterized each loop prediction as being either correct or incorrect. In turn each incorrect prediction is classified as an energy error, a sampling error, or a marginal error. Table 2 reports the total number of errors and the number of energy and sampling errors and the mean and median rmsd of the predictions from the X-ray structure.

The results in Table 2 for the 9-residue loops demonstrate that the total number of prediction errors (energy and sampling) is the lowest for the AGB implicit solvent models.

The distance-dependent dielectric model (ddd) performs the worst, followed by SGB/NP in the absence of crystal symmetry. The introduction of crystal symmetry results in a significant reduction in the number of sampling errors. (This is discussed further below.) Of the three AGB-based models, AGB-γ which mimics GB/SA is the one with the largest number of prediction errors, whereas AGBNP and AGBNP+ are equivalent in this respect. The number of energy errors, a measure of the quality of the energy model, varies greatly from one energy model to another. The fewest energy errors are found with AGBNP+, followed by in order AGBNP, AGB-γ, SGB/NP with crystal symmetry, SGB/NP, and distance-dependent dielectric. The number of sampling errors in general does not vary as greatly from one energy model to another, and their occurrence decreases significantly by using the extended sampling procedure (as shown in Table 2). This is particularly noticeable for the 13-residue loops for which two-thirds of the sampling errors with standard sampling are avoided (decrease 14 errors to five) when using extended sampling.

Comparison of the results for SGB/NP with and without crystal symmetry reveals that the inclusion of crystal symmetry has a dramatic effect on the number of sampling errors when using standard sampling; SGB/NP without crystal symmetry produces 14 sampling errors compared to five sampling errors with crystal symmetry (see Table 2). The effect of crystal symmetry on the number of sampling errors is greatly diminished when using extended sampling (Table 2). With extended sampling the number of SGB/NP sampling errors drops to seven, whereas the number of sampling errors (five) with SGB/NP with crystal symmetry is unchanged.

Table 2 also reports the mean and median rmsd of the loop predictions with respect to the X-ray structure. The mean rmsd of the 9-residue loops predictions with the AGB-based energy models is around 1 Å, which is significantly better than all the other solvation models including SGB/NP with the inclusion of crystal symmetry. The worst mean rmsd for the 9-residue loops is 2.31 Å obtained with the distance-dependent dielectric model. The median rmsd's, which are less affected by outliers corresponding to grossly incorrect predictions, are significantly smaller than the mean rmsd's. The difference between mean and median rmsd's is larger for SGB/NP-based and distance-dependent dielectric models than AGB-based solvation models due to the fact that incorrect predictions with the latter are generally closer to the X-ray structures than with the other models. The larger difference between mean and median rmsd for the 13-residue loop predictions with AGBNP+ relative to the 9-residue loop predictions reflects the fact that, expectedly, incorrect predictions with the longer loops tend to be farther away from the X-ray structure in terms of rmsd.

We repeated loop prediction calculations for six of the 9-residue protein loops classified as sampling errors with the loop prediction algorithm and using the AGBNP+ solvation model, using the T-REMD sampling procedure described in the Methods section. These loops are 1npk (residues 102−110), 1onc (70−78), 1fus (31−39), 1byb (246−254), 1noa (99−107), and 1wer (942−950) (see Table 1). We sampled these loops using temperature replica exchange molecular dynamics (T-REMD) as implemented in the IMPACT molecular mechanics package. The distribution of conformations in terms of potential energy and rmsd from the X-ray structure from the last 5 ns of the T-REMD trajectories for 1fus (31−39) is shown in Figure 4. The rmsd from the native of the lowest-energy conformations extracted from the T-REMD trajectories is reported in Table 3. For comparison, this table also reports the corresponding predictions using the standard conformational search procedure with PLOP. This table shows that in half of the cases examined (1onc, 1fus, and 1wer), T-REMD is able to produce predictions significantly closer to the X-ray structure than the PLOP-based standard sampling procedure. However, only one (1wer) of the six incorrect PLOP-based predictions results in a correct prediction with T-REMD, based on the 1.5 Å rmsd threshold value.

4. Discussion

4.1. Prediction Accuracy. The loop prediction procedure based on PLOP with the AGBNP+ solvation model and the extended sampling schemes we devised is very successful in predicting the conformations of the 9- and 13-residue loops we have investigated. As Table 2 shows, the successful prediction rate is 86% and 77% for 9- and 13-residue loops, respectively. We obtained a signficant reduction in the rates of successful predictions when using the SGB/NP and distance-dependent dielectric solvation models, even when we include crystal symmetry.

Although in this work we define the predicted conformation as the lowest-energy loop conformation, it is interesting to examine also how well the loop prediction procedure captures nativelike conformations within a given energy range from the minimum energy conformation found. In homology modeling, the choice of the candidate structures may not be restricted to selecting only the lowest-energy conformation. It may be desirable to investigate structures whose energies lie within some range about the minimum energy structure found in the search. For instance, a modeler may consider all those structures whose energies are within the lowest 5 kcal/mol as possible candidates to represent the native conformation. Under this scenario the prediction calculation can be considered successful if any one of the candidate conformations approximates well the native conformation. While the energy range is increased, the probability of including a nativelike conformation increases at the expense of the greater cost associated with having to carry over a larger number of candidate conformations. On average there are roughly 150 loop predictions per protein within 5 kcal/mol from the minimum energy. Figure 1 illustrates this cost/benefit analysis for the 57 9-residue loop prediction calculations (Table 2). Each point on the curves in Figure 1 was obtained by collecting for each loop target the set of predicted conformations with energies within a given energy range ΔE from the energy of the lowest-energy prediction and recording their number N as well as whether at least one native conformation (within 1.5 Å rmsd from the X-ray conformation) is contained in this set, that is

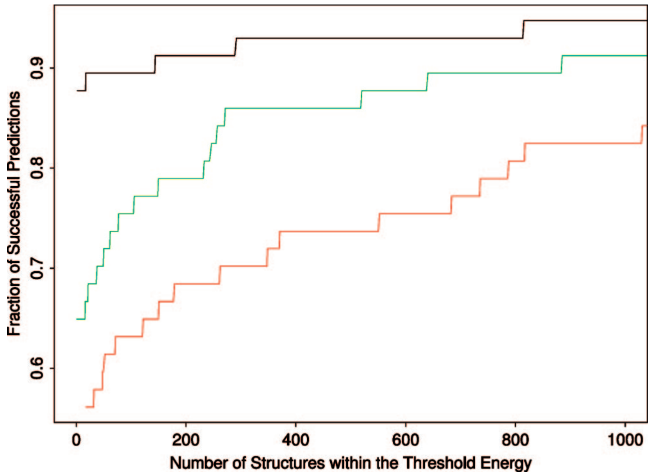


Figure 1. We plotted the ratio between the number of successfully predicted loop targets and the total number (57) of loop targets (the fraction of successful predictions) for a given threshold energy, ΔE, versus the average number of low-energy predicted conformations within this value of ΔE for the AGBNP+, SGB/NP, and distance-dependent dielectric solvation models. All loop predictions are ordered relative to their energy from the lowest-energy prediction. For an average given number of loops from the minimum (the abscissa), the fraction of proteins that have at least one nativelike loop among the top number of loops is shown above along the ordinate. The black line presents the results for AGBNP+, the green line presents the results for SGB/NP, and the red line presents the results for distance-dependent dielectric.

whether for this particular loop target and energy range the result is regarded as a successful prediction. We did this over a range of ΔE values for all 9-residue targets and solvation models. We then plotted the ratio between the number of successfully predicted loop targets and the total number (57) of loop targets (the fraction of successful predictions) for a given ΔE versus the average number of low-energy predicted conformations within this value of ΔE for the AGBNP+, SGB/NP, and distance-dependent dielectric solvation models (see Figure 1). The abscissa in this plot represents the cost, as measured by the number of conformations that one is willing to consider as possible candidates, whereas the ordinate represents the benefit, as measured by the probability of including at least one native conformation within this set of conformations. This plot can be used in two complementary ways. Given the maximum cost one is willing to sustain on the abscissa the corresponding ordinate of the curves yields for each solvation model the expected rate of success. Alternatively, given the desired rate of success in the ordinate, the curves give the required associated cost.

The minimum cost corresponds to retaining only the lowest-energy prediction ($N = 1$). This assumes that the lowest-energy loop prediction from the algorithm is the native conformation without any additional analysis. For this value of N the success rates are 86%, 77%, and 55% for the AGBNP+, SGB/NP, and distance-dependent dielectric models, respectively, see Figure 1. For all values of ΔE examined, the AGBNP+ solvation model provides the best success rate for a given cost level, followed by SGB/NP and the distance-dependent dielectric solvation models. A greater cost level entails retaining more than one low-energy loop conformation which would have to be analyzed in more detail. Conversely, AGBNP+ yields a higher success rate with less cost than the other solvation models; for example, to obtain with the SGB/NP model a success rate of 86% requires considering on average 500 conformations. To obtain a similar success with distance-dependent dielectric would require consideration of over 1000 conformations on average per loop target.

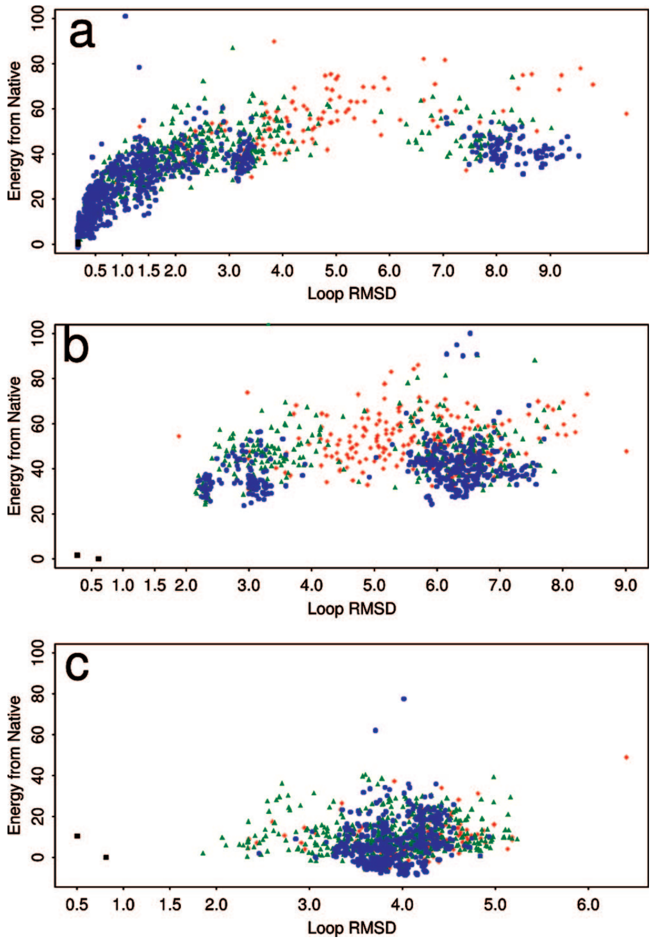


Figure 2. Energy gaps relative to the optimized native conformation (in kcal/mol) versus the rmsd (in Å) relative to the X-ray crystal conformation for three representative 9-residue loop prediction cases with the OPLS-AA/AGBNP+ potential and the standard conformational sampling algorithm: (a) 1php(91−99) (a successful prediction), (b) 1fus(31−39) (a sampling error), and (c) 3pte(215−223) (an energy error). The initial prediction results are in red, the first stage of refinement is in green, and the second stage of refinement is in blue. The native (minimized and optimized) are in black.

It is useful to compare our results with those obtained by other groups for 9-residue and 13-residue loops. Fiser et al. used MD along with simulated annealing to predict loop conformations with an all-atom force field and a statistical treatment of solvation.[24] The percentage of predictions they report within 2 Å rmsd (described as good and medium predictions) is 55%.[24] Using a tighter rmsd cutoff of 1.5 Å, we obtain with PLOP and AGBNP+ an 86% success rate in our predictions for 9-residue loops. For a set of 13-residue loops, Fiser et al., using the same 2 Å rmsd cutoff, report a very low 15% success rate,[24] compared to the 77% success rate we obtained using the AGBNP+ scoring function. Xiang et al. performed a search over a discrete rotamer library with scoring based on their colony energy. For 9-residue loops, they report an average rmsd of 2.68 Å.[25] In comparison the average rmsd we have obtained with PLOP and AGBNP+ is 1.00 Å. De Bakker et al.[62] generated loop conformations with their program RAPPER[63] and scored them with a knowledge-based potential and with a physics-based potential, AMBER/GBSA. For 9-residue loops from the Fiser set,[24] the average rmsd of the lowest-energy loops was over 2 Å when scored with the AMBER/GBSA potential which produced their best results.[62]

Jacobson et al.[26] performed loop prediction calculations on a large set of 9-residue loops using the SGB/NP model with the crystal symmetry included and using the standard conformational sampling algorithm used here.[26] Based on the Supporting Information they provided,[26] we were able to determine the number of energy and sampling errors using a 1.5 Å rmsd cutoff and a −5 kcal/mol energy cutoff. Based on our analysis of their data, they had obtained ten energy errors and eight sampling errors.[26] In comparison, we find 11 energy and seven sampling errors with SGB/NP without crystal symmetry, but we find only eight energy errors and five sampling errors with SGB/NP with crystal symmetry. This might indicate that crystal symmetry is important for prediction accuracy; however, we obtained two energy errors and five sampling errors using AGBNP+ without the presence of the crystal environment. A recent study based on the comparison of X-ray and NMR structures of identical proteins suggests that in most cases the impact of the crystal environment on protein structures is relatively small and not strongly correlated with crystal packing.[64] Recently, Zhu et al.[53,65] have reported loop prediction results for the same 35 13-residue loops investigated here using the SGB/NP potential with crystal symmetry supplemented by hydrophobic correction terms and a variable dielectric model. Zhu et al. show that these promising models lower the average backbone rmsds of the 13-residue predictions substantially, from 2.73 Å to 1.08 Å. In comparison, we obtain for the 13-residue loop set with AGBNP+ without crystal symmetry an average rmsd of 1.87 Å which is intermediate between the range of rmsd measures reported by Zhu et al.[53,65] The best performing model reported by Zhu et al. produces according to our definition five energy errors on the 13-residue loop set (see the Supporting Information of reference 65) compared with the two energy errors obtained here.

4.2. Accuracy of Scoring Functions. The ability of the effective potential model to consistently score native conformations more favorably than non-native conformations is essential for successful loop prediction. The results in Table 2 for the 9-residue loops indicate that significant differences, in terms of the number of energy errors, exist between the different solvation models we investigated. We observed that the occurrence of energy errors for each solvation model only depends weakly on the choice of conformational sampling as shown in Table 2. This is further confirmation that the energy errors are incorrect predictions mainly attributable to deficiencies of the energy functions, and as such they provide a means to analyze solvation models and suggest possible routes for improving them.

A more direct test of the potential energy functions used in loop prediction is to look at the relative percentage of energy errors rather than the relative percentage of correct predictions discussed previously which includes the effects of sampling errors. For the 9-residue loops in the absence of crystal symmetry, the largest percentage of energy errors (33.3%) was obtained for the distance-dependent dielectric. For the other implicit solvent models we tested in the absence of crystal symmetry, the percentage of energy errors decreases with, in order, SGB/NP (19.3%), AGB-γ (10.5%), AGBNP (7.0%), and AGBNP+ (3.5%).

The distance-dependent solvation model is clearly the worst in terms of accuracy, with nearly two-thirds of the incorrect predictions with extended sampling caused by energy errors (Table 2). Distance-dependent solvation models lack hydration free energy terms which provide the driving force toward solvent exposure of polar groups and vice versa the burial of hydrophobic groups. We have observed that a major structural problem with distance-dependent dielectric predictions is the occurrence of non-native salt bridges. Indeed after rescoring the distance-dependent dielectric predictions with AGBNP+, all are found to have energies greater than the native conformation due to the fact that Coulomb interaction energies of non-native ion pairs are countered by unfavorable electrostatic and nonpolar desolvation self-energy terms.

We observe about half as many energy errors with the SGB/NP solvation model as with the distance-dependent dielectric. However the occurrence of energy errors remains high; about half of the 21 incorrect predictions of 9-residue loops with SGB/NP in solution with extended sampling are attributed to the energy function. The reduction in the number of energy errors (11 to eight) with the inclusion of crystal symmetry can in principle be rationalized by the stabilization of the experimental structure due to crystal contacts not considered when evaluating the energy in solution, but we found very few examples (see below). In general the influence of the crystal environment appears to be secondary at this resolution in light of the fact that the occurrence of energy errors is significantly more pronounced with SGB/NP with crystal symmetry than with AGB-based solvation models without crystal symmetry (see Table 2). The reduction of SGB/NP energy errors with crystal symmetry is mainly due to crystal packing steric interactions preventing the formation of non-native low-energy conformations that occur in the absence of crystal contacts. Some examples illustrating the influence of the crystal environment on the loop conformation are discussed below.

Most SGB/NP predictions classified as energy errors were found to have electrostatic interaction energies significantly more negative than native conformations (results not shown), suggesting that SGB/NP overestimates the occurrence of salt bridges and intramolecular hydrogen bonds. When SGB/NP predictions are rescored with AGBNP+, all but two of the SGB/NP's energy errors are removed. Zhu et al.[53,65] recently obtained results indicating that the occurrence of energy errors with SGB/NP can be further reduced by including an empirical hydrophobic potential and a variable dielectric

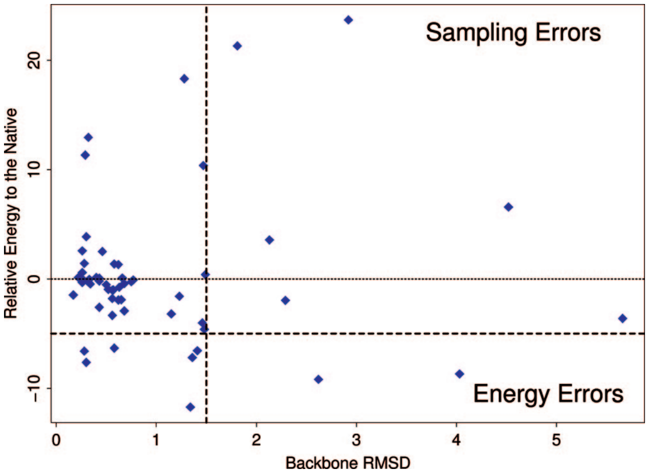


Figure 3. The results of the OPLS-AA/AGBNP+ loop predictions on the 57 9-residue loops in Table 1. The energies (in kcal/mol) relative to the native are plotted with respect to the backbone rmsd (in Å) to the native. The vertical dashed line is the rmsd cutoff, 1.5 Å. The bold, horizontal dotted-dashed line is the energy cutoff, −5 kcal/mol. Cases corresponding to the points to the left of the rmsd cutoff line are successful predictions, those in the top-right quadrant are sampling errors, and those in the bottom-right quadrant are energy errors.

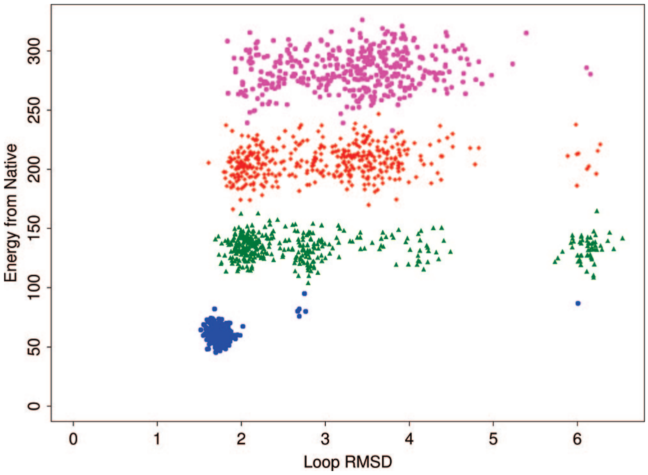


Figure 4. Energy gaps relative to the optimized native conformation (in kcal/mol) versus the rmsd (in Å) relative to the X-ray crystal conformation for the T-REMD prediction calculation of the 1fus (31−39) loop. The conformationas from the ensembles at 270 K, 400 K, 595 K, and 800 K are shown in blue, green, red, and magenta, respectively. Energies are in kcal/mol and rmsd is in Å.

model designed to favor conformations with packed hydrophobic cores and to disfavor the occurrence of salt bridges.

The AGBNP+ implicit solvent model with OPLS-AA yields only two energy errors for the 57 9-residue loops, the fewest among the solvation models tested (Table 2). The distribution of AGBNP+ results for 9-residue loops are plotted in Figure 3, where the energy errors are shown in the lower right of the plot. Only two of the 35 13-residue loop predictions with AGBNP+ are classified as energy errors. By analyzing the energy errors obtained with the various AGB-based models we are able to establish which features of the model aid in loop prediction. The number of energy errors for the 9-residue loops decreases consistently from six with the AGB-γ model, which is based on the simple surface area-only nonpolar model, to four with the AGBNP model,[44] which implements a nonpolar model that takes into account dispersive solute−solvent van der Waals interactions, to only two with the AGBNP+ model, which additionally adopts a parametrization designed to reduce the occurrence of salt bridges (see Methods).

These results indicate that the AGBNP model performs well for loop prediction applications regardless of the specific parametrization. Fine-tuning of the nonpolar model and salt bridge correction can yield, nevertheless, additional improvements. Two of the six energy errors with AGB-γ are removed when considering the AGBNP model, and, of the remaining four energy errors, two are removed when adopting ion pairing corrections in AGBNP+. One of these is the 1ivd(244−252) AGBNP prediction, which has an energy of −12.10 kcal/mol and an rmsd of 1.91 Å relative to the native. This incorrect prediction is stabilized by electrostatic interactions between Asp251 and Arg253. This interaction is absent in the 1.36 Å rmsd predicted conformation with AGBNP+, consistent with the fact that the energy of the incorrect prediction is raised above that of the correct prediction when rescored with AGBNP+. Similarly, the AGBNP incorrect prediction for 1sgp(109−117) is stabilized by a non-native ion-pair between residue Lys115 on the loop and the C-terminal carboxyl group of residue 242 which is avoided when using AGBNP+.

With AGBNP+ only two of the 13-residue loop predictions are classified as energy errors, moreover, as discussed below, the native conformations of these two loops are likely affected by intermolecular interactions present in the crystal that were not taken into account in the present calculations. In comparison, 13 of the 35 loops in this set were found to produce energy errors with the OPLS-AA/SGB/NP potential, and six of the loops are energy errors with the OPLS-AA/SGB/NP potential augmented by a hydrophobic contact correction term,[53] even though these calculations took into account crystallographic intermolecular interactions. The OPLS-AA/AGBNP+ potential function is in general able to identify the native conformation without the additional aid of knowledge-based empirical corrections, suggesting that the AGBNP solvation model captures the appropriate balance between polar and hydrophobic solvation and intramolecular interactions.

The small number of energy errors with the OPLS-AA/AGBNP+ force field are generally not very informative in terms of how to modify the potential in order to avoid them. The energy errors correspond to the 1xif(59−67) and 3pte(215−223) 9-residue loops and the 1hnj(A:191−203) and 1jp4(A:153−165) 13-residue loops. In all of these cases the native conformation is influenced by crystal contacts. Although we modeled 1xif as a monomer as did Fiser et al.[24] and Jacobson et al.,[26] the asymmetric unit of 1xif is a tetramer. However our attempt to model 1xif as a tetramer still resulted in an energy error possibly due to a native salt bridge not correctly modeled by AGBNP+. The native conformations of 3pte(215−223), 1hnj(A:191−203), and 1jp4(A:153−165) are clearly influenced by crystal packing

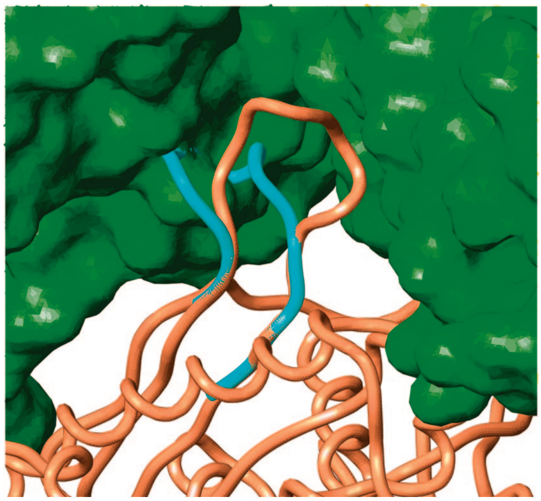


Figure 5. The X-ray (gold) and predicted (blue) conformations of the 13-residue loop 1jp4 (A:153−165). The surfaces of the crystallographically symmetric protein molecules are shown in green.

forces. As for example shown in Figure 5 for 1jp4, these loops extend away from the body of the protein, assuming a conformation unlikely to occur in solution. These loops make however extensive contacts with surrounding protein molecules in the crystal. The AGBNP+ predicted conformations without crystal symmetry instead pack closely against the protein body in a way which would not occur in the crystal due to steric repulsion. Moreover in the case of 1hnj and 1jp4, PLOP rejects backbone conformations that stray more than a certain distance from the protein body and prevents the evaluation of conformations near the native conformations. It should also be noted that, whereas we modeled only the monomer, the biological unit of 1hnj is a dimer and the loop in question (A191-A203) of one of the monomers makes contact with the same loop in the other monomer.

Apart from these cases, it appears that, within the resolution threshold we considered, the loop conformations predicted without using crystal symmetry are very close to the conformations seen in the crystal environment. This suggests that instead of crystal packing influencing loop conformations, in most cases it is the conformational propensity of the loop in solution which determines the packing arrangement in the crystal. This observation rationalizes the use of X-ray crystallographically determined structures as training sets in the development of homology modeling techniques for modeling protein loops in the solution environment.

4.3. Sampling Efficiency. Although they are indirectly influenced by properties of the energy function, such as its roughness and the level of degeneracy of native and non-native conformations, incorrect predictions classified as sampling errors primarily reflect limitations of the loop prediction algorithm. These are cases in which an incorrect prediction was made even though the energy of the native conformation is lower than that of the predicted conformation. It is important to reduce as much as possible the occurrence of sampling errors in order to decrease the overall number of mispredictions.

With the standard loop sampling procedure (Table 2) sampling errors generally represent a large fraction of incorrect predictions. This is in contrast to our results with the inclusion of crystal symmetry with the SGB/NP model in which only one-third of the incorrect predictions are classified as sampling errors. We conclude therefore that, although the parameters of the standard sampling algorithm (the value of the *ofac* parameter, the number of clusters, and the number of conformations that are passed from one stage of refinement to the next) work well when including crystallographically symmetry-related molecules,[26,53] the performance using standard sampling is significantly degraded when preforming loop prediction in the absence of the crystal environment. Evidently, the larger conformational space available to the loops in the absence of the crystal environment requires more extensive conformational search strategies. This has serious implications for loop prediction calculations as part of homology modeling projects which are typically carried out in the solution environment. Including the crystal environment is required to achieve high accuracy with the current sampling schemes. But in the majority of homology modeling applications, only the sequence and a related template protein is known. In most cases when the crystal parameters are known, so is the structure of the protein.

Sampling errors result from the sampling algorithm failing to produce near-native conformations of low enough energy or from failing to consider near-native conformations altogether. We refer to the first as a *local* sampling error and the latter as to a *global* sampling error. Global sampling errors typically occur when at the initial prediction stage the loop build-up procedure cannot find, within the resolution of the backbone and side chain rotamer library and the value of the *ofac* threshold parameter, any conformation free of clashes in the neighborhood of the native conformation. We also found that several of the global sampling errors with 9-residue loops are due to an insufficient preset number of clusters (36 for 9-residue loops), causing near-native conformations to sometimes be included in largely non-native conformational clusters. Local sampling errors are cases in which a near-native conformation produced by the initial prediction stage is abandoned prematurely and is not carried over to the subsequent refinement stages, which are responsible for adjusting the structure to lower the energy to a value closer to that of the native conformation. We found that the majority of 13-residue mispredictions are caused by local sampling errors.

Based on these observations we have modified the standard loop sampling procedure for 9-residue loops by decreasing one of the values of *ofac* tried at the initial prediction stage (from 0.75 to 0.5) and doubling the number of clusters (from 36 to 72) employed in the initial prediction stage. The standard loop sampling procedure for 13-residue loops was modified by increasing the number of candidate conformations carried over from one stage of refinement to the next (see Methods). These extended sampling schemes were then evaluated by applying them to the loops that resulted in sampling errors with the standard loop procedure. As Table 2 shows, the number of sampling errors was substantially reduced for both the 13-residue and the 9-residue loops by using the extended sampling scheme. Interestingly, none of the sampling errors obtained with SGB/NP including crystal symmetry using the standard sampling scheme improved with the extended sampling scheme, confirming the results of earlier studies[26,53] that concluded that the standard sampling procedures were sufficient for loop predictions in the crystal environment.

4.4. Loop Prediction with Replica Exchange Molecular Dynamics. To better understand the origin of the observed sampling errors we investigated with T-REMD the six 9-residue loops that resulted in global sampling errors with the standard loop sampling procedure. As has been demonstrated,[26,53] the conformational search algorithms based on PLOP perform well for predicting the conformation of protein loops of up to 13 residues in length; however, because of the exponential explosion in the number of possible loop configurations that need to be examined, the application of this method to longer loops and situations which involve several interacting loops as well as simultaneous refinement of the protein region surrounding the loops is problematic. In contrast, importance sampling schemes concentrate sampling in the most thermodynamically relevant regions of the conformational space and scale linearly with the increase of the number of degrees of freedom.

The all-atom potential energy landscapes of proteins are rugged, containing many local minima separated from each other by high barriers. Because of this there are long dwell times in local minima which slows sampling rates making application of conventional room temperature MC or MD methods impractical for loop structure determination. New simulation strategies, called collectively generalized ensemble methods,[66] have been developed which overcome this sampling bottleneck. One of the most popular methods in this class is the temperature Replica Exchange Method (REM),[66,67] which can be paired with a constant temperature molecular dynamics engine (T-REMD).[55,56,68–70] The REM technique has been used to improve sampling of rough energy landscapes. The REM methodology has been used to predict the hypervariable regions of a llama VHH antibody domain[71] and has shown promise in other protein structure determination applications.[72–74]

Prior to applying the T-REMD procedure to the group of protein loops classified as sampling errors by the standard loop prediction routine, we tested the T-REMD protocol on a less challenging set of five 9-residue loops for which the PLOP conformational search scheme was able to locate near native conformations. The T-REMD approach produced matching results within reasonable simulation times, indicating that the T-REMD protocol can also easily provide good predictions in these cases. However, as the results summarized in Table 3 show, the more challenging cases of conformational sampling, although improved over the PLOP predictions, remain problematic. The T-REMD scheme was able to substantially improve within the allocated simulation time half of the PLOP sampling errors, resulting in much higher quality structures. The rmsds of the predictions for the 1onc, 1fus, and 1wer, improved from the range between 4 Å to 7.5 Å to ~2 Å or less. Only one case, however (1wer), resulted in a correct prediction based on the 1.5 Å rmsd threshold.

The T-REMD trajectory for the 1fus (31−39) loop is illustrated in Figure 4, where the energies of conformations sampled in the last 5 ns of simulation at various temperatures are plotted. The patchy pattern of the lowest temperature ensemble of loop configurations signifies the presence of high energy barriers which separate loop configurations into different conformational states. The absence of a direct path between these structurally distinct macrostates clearly shows that efficient sampling of the conformational space would not be possible with standard molecular dynamics conducted at room temperature. Transitions between the macrostates are accomplished by acquiring enough thermal energy (moving up the temperature ladder) to surmount the separating barrier. Afterward, there is a subsequent gradual annealing of the structure and temperature leading to the native conformation at low temperature. The numbers of transitions between macrostates during 5 ns is small.

5. Conclusion

We have conducted loop conformation prediction tests on challenging benchmark sets consisting of 9- and 13-residue loops using the conformational search schemes built into PLOP to investigate the accuracy of the AGBNP implicit solvation model in conjuction with the OPLS-AA intramolecular force field. For a set of 57 9-residue loops investigated previously[24–26] we accurately predicted 88% of the loops using the OPLS-AA/AGBNP+ potential. This is a substantial improvement over the use of a distance-dependent dielectric model (63%) or SGB/NP, with (77%) or without (67%) the inclusion of crystal symmetry, as the implicit solvent model. A more substantial difference between implicit solvent models is apparent when examining the relative percentage of energy errors. AGBNP+ has the lowest percentage of energy errors at 3.5%, which is less than one-fifth as many as for SGB/NP (19.3%) and one-ninth as many as for distance-dependent dielectric (33.3%).

The fact that we have obtained high accuracy without crystal symmetry when using AGBNP+ suggests that the presence of crystal symmetry in the model is not crucial for reproducing the loop structures which have been experimentally determined via X-ray crystallography. In general, although the side chain positions have been reported to be strongly influenced by the neighboring crystallographically symmetry-related molecules,[27] the backbone conformation does not appear to be as strongly influenced by crystal packing interactions at the resolution of the current study. A recent comparison between structures determined by X-ray crystallography and NMR of identical proteins showed little correlation between structural differences and crystal contacts.[64] We found, however, the conformation sampling schemes previously developed for loop predictions in the crystal environment needed to be extended in order to avoid sampling errors when crystal symmetry is not included in the model. We recommend the use of these updated extended sampling protocols for homology modeling applications in the solution environment.

We expect importance sampling conformational search methods such as T-REMD to become an important complement to traditional discrete conformational search methods in cases when the number of degrees of freedom is large such as interacting loops, imperfect frameworks for loop prediction, etc. We note that development of better implementations of REM ideas which will offer faster sampling in the context of structure prediction of protein loops is the subject of intensive ongoing research. This will go beyond simple temperature exchanges in REM and will involve modifying the system Hamiltonians and swapping replicas with different energy potentials, constructed to effectively increase the range of conformational motion.[71] Another avenue of improvement is to consider more rational ways of selecting pairs of replicas for exchanges of temperatures or Hamiltonian parameters,[75] with the goal being to examine how sampling can be enhanced through maximizing mixing among replicas. Such a multidimensional replica exchange procedure appears to be promising for exploring the conformational space of protein loops.

It should be noted that the success rates we obtained likely overestimate the success rate obtainable in actual homology modeling applications because these tests were performed in the idealized case in which the frame of the protein surrounding the loop is known. Successful prediction in this idealized situation is a necessary but not sufficient requirement for the ability to predict the correct nativelike loop conformation with partial knowledge of the protein framework. We have begun to investigate cases in which the conformations of the protein side chains surrounding the loop are predicted at the same time as the loop conformation. We find that the successful prediction rate for these cases is significantly reduced relative to the tests reported here with the conformations of the side chains of the protein frame fixed in their native conformations. Clearly more work is still needed to develop fast and accurate loop prediction protocols for “real life” homology modeling applications.

Acknowledgment. This project has been supported in part by the National Institutes of Health Grants (GM-30580 to R.M.L. and GM-52018 to R.A.F.). We thank Matt Jacobson for providing the PLOP program and Kai Liu for providing the script on which we based our 13-residue prediction protocol. We also thank Jennifer Knight for assistance with implementing PLOP and predicting loops with imperfect frames.

References

(1) Ginalski, K. *Curr. Opin. Struct. Biol.* **2006**, *16*, 172–177.

(2) Kryshtafovych, A.; Venclovas, C.; Fidelis, K.; Moult, J. *Proteins* **2005**, *61*, 225–236.

(3) Shiffer, C.; Hermans, J. *Methods Enzymol.* **2003**, *374*, 412–461.

(4) Xia, B.; Tsui, V.; Case, D.; Dyson, H.; Wright, P. *J. Biomol. NMR* **2002**, *22*, 317–331.

(5) Skolnick, J. *Curr. Opin. Struct. Biol.* **2006**, *16*, 166–171.

(6) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145.

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

(8) Huang, E. S.; Subbiah, S.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1996**, *257*, 716–725.

(9) Park, B.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 367–392.

(10) Park, B. H.; Huang, E. S.; Levitt, M. *J. Mol. Biol.* **1997**, *266*, 831–846.

(11) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399–1401.

(12) Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. *Proteins: Struct., Funct., Genet.* **1999**, *S3*, 171–176.

(13) Lazaridis, T.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 477–487.

(14) Petrey, D.; Honig, B. *Protein Sci.* **2000**, *9*, 2181–2191.

(15) Bursulaya, B. D.; Brooks III, C. L. *J. Phys. Chem. B* **2000**, *104*, 12378–12383.

(16) Dominy, B. N.; Brooks, C. L. *J. Comput. Chem.* **2002**, *23*, 147–160.

(17) Liu, Y.; Beveridge, D. L. *Proteins: Struct. Funct. Genet.* **2002**, *46*, 128–146.

(18) Feig, M.; Brooks, C. L., III *Proteins: Struct. Funct. Genet.* **2002**, *49*, 232–245.

(19) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 404–422.

(20) Zhang, Y.; Kolinski, A.; Skolnick, J. *Biophys. J.* **2003**, *85*, 1145–1164.

(21) Tsai, J.; Bonneau, R.; Morozov, A.; Kuhlman, B.; Rohl, C.; Baker, D. *Proteins* **2003**, *53*, 76–87.

(22) Wang, K.; Fain, B.; Levitt, M.; Samudrala, R. *BMC Struct. Biol.* **2004**, *4*, 8.

(23) Qiu, J.; Elber, E. *Proteins* **2005**, *61*, 44–55.

(24) Fiser, A.; Do, R. K. G.; Sali, A. *Protein Sci.* **2000**, *9*, 1753–1773.

(25) Xiang, Z. X.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432–7437.

(26) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins: Struct., Funct., Bioinform.* **2004**, *55*, 351–367.

(27) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597–608.

(28) Sherman, W.; Day, T.; Jacobson, M.; Friesner, R.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.

(29) Levy, R. M.; Gallicchio, E. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531–67.

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

(31) Dominy, B. N.; Brooks III, C. L. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(32) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(33) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591–1608.

(34) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128–137.

(35) Banks, J.; et al., *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(36) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

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

(38) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275–291.

(39) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(40) Lee, M. S.; Feig, M., Jr.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(41) Feig, M.; Onufriev, A.; Lee, M.; Im, W.; Case, D.; Brooks, C., III *J. Comput. Chem.* **2004**, *25*, 265–284.

(42) Zhang, L.; Gallicchio, E.; Friesner, R.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591–607.

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

(44) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

(45) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(46) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(47) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(48) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

(49) Levy, R. M.; Zhang, L. Y.; Gallicchio, E. amd Felts, A. K. *J. Am. Chem. Soc.* **2003**, *25*, 9523–9530.

(50) Su, Y.; Gallicchio, E. *Biophys. Chem.* **2004**, *109*, 251–260.

(51) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(52) Gallicchio, E.; Zhang, L.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(53) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. *Proteins: Struct., Funct., Bioinform.* **2006**, *65*, 438–452.

(54) Hartigan, J. A.; Wong, M. A. *Appl. Stat.* **1979**, *28*, 100–108.

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

(56) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinform.* **2004**, *56*, 310–321.

(57) Banks, J. L. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

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

(59) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.

(60) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory. Comput.* **2006**, *2*, 115–127.

(61) Wang, G.; Dunbrack, R. *Bioinformatics* **2003**, *19*, 1589–1591.

(62) de Bakker, P. I. W.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins: Struct., Funct., Bioinform.* **2003**, *51*, 21–40.

(63) DePristo, M. A.; de Bakker, P. I. W.; Lovell, S. C.; Blundell, T. L. *Proteins: Struct., Funct., Bioinform.* **2003**, *51*, 44–55.

(64) Andrec, M; Snyder, D. A.; Zhou, Z.; Young, J. T. M. G.; Levy, R. M. *Proteins: Struct., Funct., Bioinform.* **2007**, *69*, 449–465.

(65) Zhu, K.; Shirts, M. R.; Friesner, R. A. *J. Chem. Theory Comput.* **2007**, *3*, 2108–2119.

(66) Sugita, Y.; Okamoto, Y. Free-energy calculations in protein folding by generalized-ensemble algorithms. In *Lecture Notes in Computational Science and Engineering*; Schlick, T.; Gan, H. H., Eds.; Springer-Verlag: Berlin, 2002.

(67) Nymeyer, H.; Gnanakaran, S.; García, A. E. *Methods Enzymol.* **2003**, *383*, 119–149.

(68) García, A. E.; Sanbonmatsu, K. Y. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 345–354.

(69) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–14936.

(70) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. *J. Chem. Phys.* **2004**, *121*, 10748–10756.

(71) Fenwick, M. K.; Escobedo, F. A. *Biopolymers* **2003**, *68*, 160–177.

(72) Chen, J.; Im, W.; Brooks III, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 16038–16047.

(73) Habeck, M.; Nilges, M.; Rieping, W. *Phys. Rev. Lett.* **2005**, *94*, 018105.

(74) Nanias, M.; Chinchio, M.; Oldziej, S.; Czaplewski, C.; Scheraga, H. A. *J. Comput. Chem.* **2005**, *26*, 1472–1486.

(75) Calvo, F. *J. Chem. Phys.* **2005**, *123*, 124106.

CT800051K