JCTC Journal of Chemical Theory and Computation

Assessment and Validation of the Electrostatically Embedded Many-Body Expansion for Metal−Ligand Bonding

Duy Hua, Hannah R. Leverentz,* Elizabeth A. Amin, and Donald G. Truhlar

Department of Chemistry, Department of Medicinal Chemistry, and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455, United States

Received August 27, 2010

Abstract: The electrostatically embedded many-body method has been very successful for calculating cohesive energies and relative conformational energies of clusters, and here we extend it to calculate bond breaking energies for metal−ligand bonds in inorganic coordination chemistry. We find that, on average, the electrostatically embedded pairwise additive method is able to predict bond energies yielded by conventional full-system calculations done at the same level of theory to within 2.5 kcal/mol and that the electrostatically embedded three-body method consistently yields energies within 1.0 kcal/mol of the full-system calculations.

An important theme of modern quantum chemistry is enabling reliable calculations on large and complex systems. A general strategy for achieving this is fragmentation, and a variety of fragmentation schemes have been explored.[1−12] One especially promising approach is the electrostatically embedded many-body[5,13−18] (EE-MB) expansion. In work reported so far, we have obtained accurate results at low cost for noncovalently bonded clusters. For example, the electrostatically embedded three-body (EE-3B) method based on CCSD(T) calculations of dimers and trimers was able to reproduce full water hexamer calculations with a mean unsigned deviation of only 0.12 kcal/mol (0.3%), whereas full CCSD calculations using the same basis set yielded a mean unsigned relative deviation of 0.43 kcal/mol (0.9%),[14] and to reproduce full calculations on (H_2SO_4)-$(HSO_4^-)(NH_4^+)(H_2O)_6$ with a mean unsigned deviation of only 0.15 kcal/mol (0.2%).[17] In the present letter, we apply these methods to bond breaking energies of complexes bound by coordinate covalent[19] bonds.

The EE-MB method is described elsewhere,[5] and so it is reviewed only briefly here. The fragments into which the system is partitioned are called monomers. We will test two variants: the electrostatically embedded pairwise additive (EE-PA) method and the EE-3B method. In EE-PA, the energy of systems composed of monomers m, n, p, etc. is approximated as

$$E^{\mathrm{PA}} = E^{(1)} + \Delta E^{(2)} \tag{1}$$

where

$$E^{(1)} = \sum_m E_m \tag{2}$$

$$\Delta E^{(2)} = \sum_m \sum_{n>m} \Delta E_{mn}^{(2)} \tag{3}$$

$$\Delta E_{mn}^{(2)} = E_{mn} - E_m - E_n \tag{4}$$

and the EE-3B energy is

$$E^{\mathrm{3B}} = E^{\mathrm{PA}} + \Delta E^{(3)} \tag{5}$$

where

$$\Delta E^{(3)} = \sum_m \sum_{n>m} \sum_{p>n} \Delta E_{mnp} \tag{6}$$

$$\Delta E_{mnp} = E_{mnp} - E_{mnp}^{\mathrm{PA}} \tag{7}$$

where E_m, E_{mn}, and E_{mnp} are the energies of a monomer, dimer, and trimer, respectively, each embedded in a field of point charges representing the other monomers, and E_{mnp}^{PA} is the EE-PA approximation to the energy of trimer mnp. The individual embedded oligomer energies (where oligomer is a general term that can be replaced by monomer, dimer, or trimer) can be computed using any desired level of electronic structure theory. Most levels of electronic structure theory require that any system on which they are used have an integer number of electrons; therefore, charge transfer between oligomers of a given type usually cannot occur within most practical applications of the EE-MB approximation (including the present study).

All calculations were carried out with the M05-2X density functional[20] and the B2 basis set,[21] which is a polarized valence triple-ζ basis set optimized for use with Zn-containing complexes. The innermost 10 electrons (small core) of Zn are replaced by the (MEFIT,*R*) relativistic effective core potential.[22,23] The M05-2X density functional was chosen because of previous tests[21,24] that showed it to yield high accuracy results for Zn-containing complexes and biological structures.

* Correponding author e-mail: lever046@umn.edu.

10.1021/ct100491q
Published on Web 12/29/2010

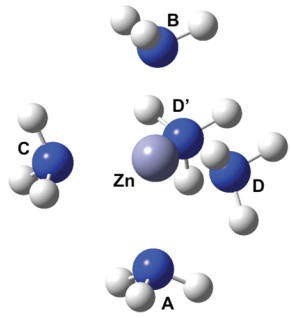


Figure 1. Structure of $Zn(NH_3)_5^{2+}$.

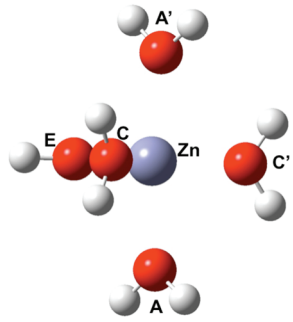


Figure 2. Structure of $Zn(H_2O)_5^{2+}$.

All conventional calculations were performed using *Gaussian 09*.[25] All EE-MB calculations were performed using MBPAC 2009-2,[26] which is a program that requires the user to define which atoms in the overall system belong to each monomer and then calls MN-GFM,[27] a locally modified version of *Gaussian 03*[28] to perform the necessary monomer, dimer, and trimer calculations.

No attempt was made to correct for basis set superposition error (BSSE)[29,30] in any of the present calculations for the following reasons: (1) The goal of this work is not to obtain high-level benchmark calculations of the binding energies of coordinate covelent clusters; rather, the goal of this work is to show that the EE-MB method can yield metal−ligand binding energies close to those of conventional calculations performed at a given level of theory. Therefore, attempting to correct for BSSE in the present study would be somewhat superfluous because it is not necessary in order to achieve our goal. (2) The counterpoise correction (CP)[29] for BSSE is only clearly defined for dimer interactions. Rigorous extensions of the CP correction for BSSE in trimer interactions and beyond have been proposed,[31] but these methods are computationally costly. Rather than investing computer time in applying a many-body extension of the CP correction, one can invest that computer time (or less) into performing a non-BSSE-corrected calculation with a basis set large enough to preclude the need for such corrections. Often, a triple-ζ basis set with a judicious choice of polarization and diffuse functions will suffice.[32]

The systems that we consider are (**1**) $Zn(NH_3)_5^{2+}$ (shown in Figure 1), (**2**) $Zn(H_2O)_5^{2+}$ (Figure 2), and (**3**) $Zn(H_2O)_4(OH)^+$ (Figure 3). The quantities we calculate are bond breaking energies defined as the energy to remove one of the ligands from the complex, with the internal coordinates of both fragments frozen; thus, the quantity calculated here is not a conventional bond dissociation energy but rather a relative energy along a cut through the potential energy surface with

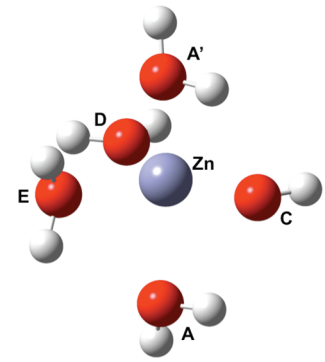


Figure 3. Structure of $Zn(H_2O)_4(OH)^+$.

fixed fragment geometries. This provides a direct test of the ability of EE-MB to predict relative electronic energies along bond breaking coordinates, with no complications from competing effects of relaxation. The energy of breaking the bond is the sum of the energies of the two products (separated frozen fragments) minus the energy of the reactant. When calculating the energies of a product, the embedding charges of the other (separated) product are not present because the other product is infinitely separated.

The geometries of the three complexes were optimized with M05-2X/B2/(MEFIT,*R*). Each of the coordination complexes has the structure of an irregular trigonal bipyramid, with axial ligands A and B and equatorial ligands C, D, and E. If r_X denotes the distance from the nonhydrogenic atom of a ligand to Zn, we label the atoms so that $r_A \leq r_B$ and $r_C \leq r_D \leq r_E$. Furthermore, if $r_B = r_A$, then B is also called A′; if $r_D = r_C$, then D is also called C′, and if $r_E = r_D$, then E is also called D′.

The partial atomic charges on the fragments are calculated in every case for the isolated fragment at the geometry it has in the entire system. For example, if we are calculating the energy of ZnABCDE, and if one of the fragments is ZnCD, we calculate the partial atomic charges of ZnCD by removing A, B, and E from the system, and if another fragment is E, we calculate its partial charges by removing Zn, A, B, C, and D. Thus, the calculations used to obtain charges are the same as the monomer calculations of eq 2 except that, for obtaining charges, the monomers are not embedded. We do not want the results to depend strongly on the method used to calculate charges, and therefore we used three different methods to calculate charges in order to test the sensitivity. The methods used are Hirshfeld population analysis (HPA),[33] natural population analysis (NPA),[34] and Merz−Kollman (MK) electrostatic fitting.[35]

For each of the three complexes, we considered two dissociation processes:

$$ZnABCDE^{n+} \rightarrow ZnABCD^{n+} + E \qquad (R1)$$

and

$$ZnABCDE^{n+} \rightarrow ZnBCDE^{n+} + A \qquad (R2)$$

where n is 2 (Figures 1 and 2) or 1 (Figure 3). Notice that process R1 breaks an equatorial bond, and process R2 breaks an axial bond.

In preliminary work, we found that taking an isolated Zn^{2+} or a metal ion with a single ligand as a fragment did not yield accurate results; this is understandable because the Hirshfeld

Table 1. Target Bond Energies (kcal/mol)

process	complex	broken bond	bond energy
1	**1**	Zn−A	34.90
2	**1**	Zn−D′	45.97
1	**2**	Zn−A	40.06
2	**2**	Zn−E	45.68
1	**3**	Zn−A	25.22
2	**3**	Zn−E	33.23

Table 2. Unsigned Errors in Bond Energies (kcal/mol)

	EE-PA			EE-3B		
process	HPA	NPA	MK	HPA	NPA	MK
1	0.10	0.59	0.75	0.78	0.14	0.20
2	2.06	3.06	2.60	0.79	0.15	0.22
3	2.21	2.27	2.04	0.65	0.91	0.81
4	3.10	2.35	2.34	0.65	0.91	0.81
5	1.80	2.24	1.98	0.09	0.44	0.26
6	2.95	2.63	2.58	0.08	0.45	0.27
mean[a]	2.04	2.19	2.05	0.51	0.50	0.42

[a] Mean unsigned error in all six cases.

partial atom charges of Zn in Zn^{2+}, $Zn(H_2O)^{2+}$, and $Zn(NH_3)^{2+}$, at typical geometries in the latter two cases, are 2, 1.6, and 1.4, respectively, whereas in the larger fragments, the Hirshfeld partial atomic charge on Zn is much lower, in the range 0.5−1.2. Thus, we only consider fragmentation schemes where one of the fragments is Zn^{2+} with two ligands. In particular, because $r_{equatorial} < r_{axial}$, we take one fragment as $ZnCD^{2+}$ for Figures 1 and 2 and $ZnCD^{+}$ for Figure 3. Note that A and B are NH_3 in Figure 1, they are H_2O in Figure 2, and A is OH^- and B is H_2O in Figure 3. We do not break the bond within the fragment because that gives a product where Zn has only one ligand in the fragment. In all cases, one fragment is Zn with the two closest ligands (which are always equatorial ligands), and the other fragments are individual ligands. We then consider breaking the bond between Zn and the farthest equatorial ligand or the bond between Zn and the nearest axial ligand.

Our goal is to test how well the EE-MB approximation can reproduce a result on the whole complex. Therefore we first performed M05-2X/B2/(MEFIT,*R*) calculations on bond breaking without using the many-body approximation; the results are in Table 1. The table shows that the bond breaking energies vary by more than 20 kcal/mol for the various test cases. If we can predict the bond energies within 1 kcal/mol, we will have captured the variation within 5%. Table 2 shows the results. EE-PA has a mean unsigned error of about 2 kcal/mol, whereas EE-3B has a mean unsigned error of 0.4−0.5 kcal/mol, with a maximum error of 0.9 kcal/mol. So we judge the EE-3B theory to be a success for near-equilibrium structures.

In order to show that the EE-PA and EE-3B approximations also yield consistent performance on higher-energy, nonequilibrium structures, we selected several points along a slice of the potential energy surface (PES) of the $[Zn(NH_3)_5]^{2+}$ system shown in Figure 1. We adjusted the distance between Zn and the nitrogen atom of ammonia molecule D′, holding the remaining $[Zn(NH_3)_4]^{2+}$ complex (composed of Zn^{2+}, A, B, C, and D) and the D′ ammonia molecule rigid in the geometries that they had in the $[Zn(NH_3)_5]^{2+}$ complex. At each Zn−N distance (*R*(Zn−N)), we calculated the conventional, EE-PA, and EE-3B M05-2X/B2/(MEFIT,*R*) energies of the system. The

Table 3. Conventional Binding Energies (BE) and Errors in EE-PA and EE-3B Binding Energies As a Function of *R*(Zn−N) (All Values in kcal/mol)

R(Zn−N) (Å)	conventional BE (kcal/mol)	EE-PA error (kcal/mol)	EE-3B error (kcal/mol)
1.721	−21.26	−0.10	−0.75
1.821	−34.51	−0.61	−0.72
1.921	−41.73	−1.23	−0.61
2.021	−45.08	−1.90	−0.43
2.121[a]	−45.97	−2.60	−0.22
2.221	−45.34	−3.27	−0.01
2.421	−41.68	−4.43	0.27
2.721	−34.32	−5.37	0.20
3.121	−25.52	−5.28	−0.43
4.121	−13.63	−2.61	−0.47
5.121	−8.66	−0.58	−0.09
7.121	−4.63	0.02	−0.01
MSE[b]		−2.33	−0.27
MUE[b]		2.33	0.35
RMSE[b]		2.98	0.43

[a] This is the equilibrium geometry of the complex. [b] MSE = mean signed error, MUE = mean unsigned error, RMSE = root mean squared error.

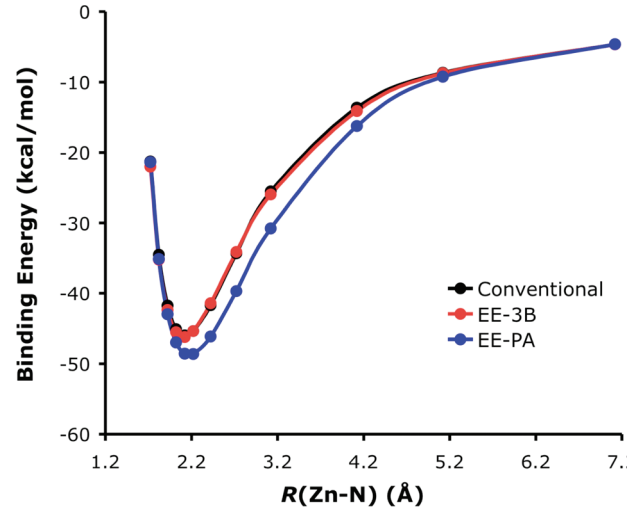


Figure 4. $Zn(NH_3)_5^{2+}$ binding energy (in kcal/mol) as a function of the distance between zinc and the nitrogen atom of the ammonia molecule D′ (*R*(Zn−N), in Å).

MK charges of the isolated rigid monomers $ZnCD^{2+}$, A, B, and D′ were used as the embedding charges in the EE-MB calculations. Table 3 shows the conventional binding energy and the signed errors in the EE-PA and EE-3B binding energies (relative to the conventional energies) at each *R*(Zn−N). Figure 4 shows the actual binding energies calculated at each point on the PES slice using the conventional, EE-PA, and EE-3B methods. We again find that EE-PA has a mean unsigned error of about 2 kcal/mol and that EE-3B has a mean unsigned error of less than 0.4 kcal/mol. Therefore, we conclude that the EE-3B method is a success for a variety of metal−ligand complex geometries.

The present study shows that the EE-3B method consistently yields bond energies within 1 kcal/mol of those obtained from the full calculation at a given level of electronic structure theory. This is encouraging because using the EE-3B method instead of a conventional calculation is advantageous for the following reasons: (a) An EE-MB calculation at any level of electronic structure theory can easily be made to run on several processors

at once in order to save wall clock time. (b) A series of small calculations is less likely to exceed the memory or disk capacity of a computer system than is a single large calculation. (c) For systems with a large number of fragments (e.g., a coordinate complex in explicit solvent), an EE-3B calculation scales more favorably with the size of a chemical system than do conventional calculations at a high level of electronic structure theory (i.e., hybrid density functional theory or correlated wave function theory).

Previous studies (cited earlier) have shown that the EE-3B method is consistently able to predict binding energies of noncovalently bonded systems to within 1 kcal/mol (and often much less) of a conventional full-system calculation performed at the same level of theory. The complexes examined in those studies were held together by hydrogen bonds, intermonomer dispersion-like interactions, and electrostatic interactions between oppositely charged ions. The present study has examined systems held together by a stronger type of interaction: the coordinate covalent bond. We found that the EE-3B method consistently yields unsigned errors of less than 1 kcal/mol (relative to conventional calculations performed at the same level of theory) in the bond breaking energies of six bond breaking processes in three different zinc-ligand complexes when the zinc ion and the two ligands closest to it are treated as a single monomer. Consistent with previous studies, the errors in the EE-3B method do not depend strongly on the charge analysis method used to obtain the embedding charges; the EE-3B method performs well regardless of which set of embedding charges is used. We conclude that the EE-3B method is a convenient and accurate way to study relative electronic energies, i.e., potential energy surfaces, along bond breaking coordinates in coordinate covalent complexes.

Acknowledgment. The authors would like to thank Dr. Richard Wood for participation in the early stages of this project. This work was supported in part by the National Science Foundation under grant no. CHE09-56776 and by the Lando/NSF summer undergraduate research program of the Department of Chemistry of the University of Minnesota.

References

(1) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.

(2) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2004**, *120*, 6832.

(3) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.

(4) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777.

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

(6) Collins., M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.

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

(8) Hirata, S.; Yagi, K. *Chem. Phys. Lett.* **2008**, *464*, 123.

(9) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2008**, *128*, 234108.

(10) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. *J. Phys. Chem. B* **2009**, *113*, 9646.

(11) Söderhjelm, P.; Aquilante, F.; Ryde, U. *J. Phys. Chem. B* **2009**, *113*, 11085.

(12) Li, W.; Piecuch, P. *J. Phys. Chem. A* **2010**, *114*, 6721.

(13) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.

(14) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.

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

(16) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.

(17) Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1573.

(18) Speetzen, E. D.; Leverentz, H. R.; Lin, H.; Truhlar, D. G. In *Accurate Condensed Phase Electronic Structure Theory*; Manby, F., Ed.; CRC Press: Boca Raton, FL, 2010.

(19) Cotton, F. A.; Wilkinson, G. *Advanced Inorganic Chemistry: A Comprehensive Text*, 3rd ed.; Interscience Publishers: New York, 1972; p 72.

(20) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241; Erratum: **2008**, *119*, 525.

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

(22) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866.

(23) Kaupp, M.; Stoll, H.; Preuss, H. *J. Comput. Chem.* **1990**, *11*, 1029.

(24) Sorkin, A.; Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1254.

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.01; Gaussian, Inc.: Wallingford, CT, 2009.

(26) Dahlke, E. E.; Lin, H.; Leverentz, H.; Truhlar, D. G. *MBPAC 2009-2*; University of Minnesota: Minneapolis, MN, 2009.

(27) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module*, version 4.1; University of Minnesota: Minneapolis, MN, 2009.

(28) 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. *Gaussian03*, Revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.

(29) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(30) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418. Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024; Errata: **1987**, *86*, 3760.

(31) Valiron, P.; Mayer, I. *Chem. Phys. Lett.* **1997**, *275*, 46.

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

(33) Hirschfeld, F. L. *Theor. Chem. Acc.* **1977**, *44*, 129.

(34) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.

(35) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.

CT100491Q

JCTC Journal of Chemical Theory and Computation

Shannon Entropy Based Time-Dependent Deterministic Sampling for Efficient "On-the-Fly" Quantum Dynamics and Electronic Structure

David Hocker,[†] Xiaohu Li,[‡] and Srinivasan S. Iyengar*

Department of Chemistry and Department of Physics, Indiana University, 800 E. Kirkwood Ave, Bloomington, Indiana 47405, United States

Received October 14, 2010

Abstract: A new set of time-dependent deterministic sampling (TDDS) measures, based on local Shannon entropy, are presented to adaptively gauge the importance of various regions on a potential energy surface and to be employed in "on-the-fly" quantum dynamics. Shannon sampling and Shannon entropy are known constructs that have been used to analyze the information content in functions: for example, time-series data and discrete data sets such as amino acid sequences in a protein structure. Here the Shannon entropy, when combined with dynamical parameters such as the instantaneous potential, gradient and wavepacket density provides a reliable probe on active regions of a quantum mechanical potential surface. Numerical benchmarks indicate that the methods proposed are highly effective in locating regions of the potential that are both classically allowed as well as those that are classically forbidden, such as regions beyond the classical turning points which may be sampled during a quantum mechanical tunneling process. The approaches described here are utilized to improve computational efficiency in two different settings: (a) It is shown that the number of potential energy calculations required to be performed during on-the-fly quantum dynamics is fewer when the Shannon entropy based sampling functions are used. (b) Shannon entropy based TDDS functions are utilized to define a new family of grid-based electronic structure basis functions that reduce the computational complexity while maintaining accuracy. The role of both results for on-the-fly quantum/classical dynamics of electrons and nuclei is discussed.

I. Introduction

The time-dependent Schrödinger equation is the starting point for many computational methodologies employed in gas-phase[1] and condensed-phase chemical dynamics.[2] When utilized, the Born−Oppenheimer approximation allows for separation of the nuclear and electronic degrees of freedom in a system, allowing for varying treatments of the nuclei, be it classical,[3−7] quantum-mechanical,[1,8−34] or semiclassical.[35−42] In all cases, the nuclei are either propagated along parametrically fitted electronic surfaces known *a priori*, or along highly accurate (and sometimes computationally expensive) electronic surfaces that require no prior knowledge of the system. Due to the large number of quantum mechanical energy and gradient calculations required by the latter approach, there has been a strong motivation toward "on-the-fly" dynamics schemes to overcome this computational barrier and potentially allow for larger, more complex systems to be studied.[3−7,35,41,43−46] This growing subfield of *ab initio* molecular dynamics (AIMD) approximates the electronic structure alongside the nuclei to simulate molecular dynamics. When AIMD techniques are embedded in a full quantum or semiclassical scheme, there is the potential for large systems to be accurately treated with the complete machinery of quantum dynamics. Several efforts have been made toward this goal.[41,46−49]

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

[†] Present address: Department of Chemistry, Princeton University.

[‡] Present address: Department of Chemistry, Northwestern University.

10.1021/ct1005856
Published on Web 01/12/2011

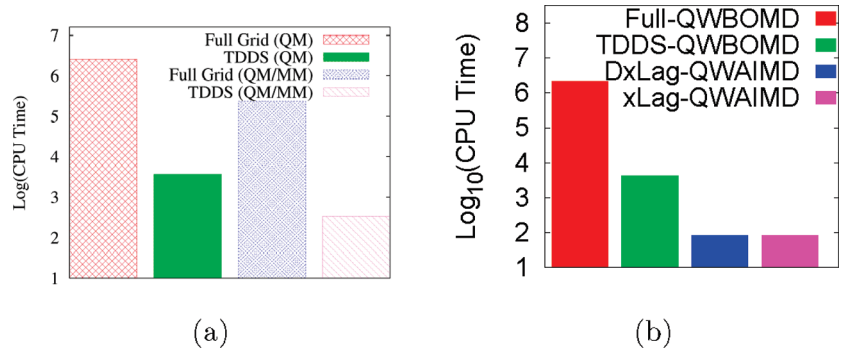


Figure 1. (a) Computational expense for QWAIMD with and without time-dependent deterministic sampling (TDDS). *Note that in all cases the vertical axis is the logarithm of CPU time.* TDDS provides an enormous reduction in computational time for two different types of embedding schemes (QM/MM and QM), with little loss in accuracy. (Reproduced with permission from ref 53. Copyright 2008, American Institute of Physics). (b) Further reduction in computation time is facilitated through the introduction of a propagation scheme that involves multiple diabatic states.[54] Again, accuracy in computing potential surfaces is preserved, while reducing the computational overhead substantially. (Reproduced with permission from ref 54. Copyright 2010. American Institute of Physics).

Recently,[46,50−56] we have introduced a methodology that accurately computes quantum dynamical effects in a subsystem while simultaneously handling the motion of the surrounding atoms and changes in electronic structure calculation. The approach is quantum-classical[40,57−63] and involves the synergy between a time-dependent quantum wave packet description and *ab initio* molecule dynamics. As a result, the approach is called quantum-wave packet *ab initio* molecular dynamics (QWAIMD). Since the quantum dynamics is performed on Cartesian grids, the predominant bottleneck is the computation of the grid-based, time-dependent electronic structure potential and gradients generated by the motion of the classical nuclei. This limitation is partially surmounted through the following methodological improvements:

(a) A time-dependent deterministic sampling (TDDS) technique was introduced in refs 51 and 52, which when combined with numerical methods such as an efficient wavelet compression scheme and low-pass filtered Lagrange interpolation,[52] provides computational gains of many orders of magnitude (Figure 1).

(b) Multiple diabatic reduced single particle electronic density matrices are propagated simultaneously with the quantum wavepacket in ref 54, and the associated diabatic states are used to construct an adiabatic surface at every instant in time using a nonorthogonal CI formalism. The diabatic approximation allows reuse of the two-electron integrals during the on-the-fly potential energy surface computation stage and leads to substantial reduction in computational costs (Figure 1).

QM/MM generalizations to QWAIMD have also been completed.[53] The approach is being generalized to treat extended systems[64] for condensed-phase simulations; a biased QWAIMD formalism to sample rare events is also currently being developed. We have utilized QWAIMD to compute vibrational properties of hydrogen-bonded clusters inclusive of quantum nuclear effects[52] and have also adopted the method to study hydrogen tunneling in enzyme active sites.[55,65] The quantum dynamics scheme in QWAIMD has also been used to develop a technique known as multistage *ab initio* wavepacket dynamics (MSAIWD) to treat open-electronic systems.[66,67]

In this publication, we probe the relevant regions of a potential surface, using a new TDDS function based on the notion of Shannon entropy.[68−72] This paper is organized as follows: An overview of QWAIMD is presented in section II along with a discussion of time-dependent deterministic sampling and its current efficacy. The derivation and physical rationale for the sampling functions that utilize Shannon entropy are given in section III. The numerical benchmarks are arranged in a multipronged fashion. In section IV.A, we discuss the use of the Shannon entropy based TDDS functions in adaptive determination of critical regions of the potential surface during dynamics. Accuracy in computing vibrational properties is also discussed. In section IV.B, the approach is utilized to construct an accurate "grid-based" electronic basis set. This implementation of Shannon-entropy based TDDS leads to a sizable reduction in the number of electronic basis functions that need to be utilized in calculations involving hydrogen-bonded systems. Consequently, the approach is tested for accuracy and efficiency for three different kinds of hydrogen-bonded clusters. This same idea is further exploited in ref 54 to develop an implicitly time-dependent, grid-based electronic structure basis to tremendously improve the efficiency and accuracy of QWAIMD. Concluding remarks are given in section V.

II. Main Features of Quantum Wavepacket Ab Initio Molecular Dynamics (QWAIMD)

As noted above, QWAIMD is based on a synergy between quantum wavepacket dynamics and *ab initio* molecular dynamics.[46,50−55,64] The partitioning scheme divides the system into three subsystems: subsystem A may include particles that display critical quantum dynamical effects; subsystems B and C contain the surrounding nuclei and electrons, respectively, and are treated under the AIMD formalism.[46,50,51,73,74] Subsystem A is propagated according to the Trotter-factorized quantum mechanical time propagator:[10,75−77]

$$\chi_A(x;t) = \exp\left\{-\frac{\imath Ht}{\hbar}\right\}\chi_A(x;t=0) = \left[\exp\left\{-\frac{\imath Vt}{2\hbar}\right\}\exp\left\{-\frac{\imath Kt}{\hbar}\right\}\exp\left\{-\frac{\imath Vt}{2\hbar}\right\} + \mathcal{O}(t^3)\right]\chi_A(x;t=0), \quad (1)$$

and the free-propagator, $\exp\{-(\imath Kt)/(\hbar)\}$, is represented using "distributed approximating functionals" (DAF):[46,50,78−81]

$$\left\langle x\left|\exp\left\{-\frac{\imath K\Delta t}{\hbar}\right\}\right|x'\right\rangle \equiv \tilde{K}(x,x';\Delta t) \equiv \tilde{K}(|x-x'|,\Delta t) = \frac{(2\pi)^{-1/2}}{\sigma(0)}[e^{-[(x-x')^2]/[2\sigma(\Delta t)^2]}] \times \sum_{n=0}^{M_{DAF}/2}\frac{(-1/4)^n}{n!}\left(\frac{\sigma(0)}{\sigma(\Delta t)}\right)^{2n+1}H_{2n}\left(\frac{x-x'}{\sqrt{2}\sigma(\Delta t)}\right) \quad (2)$$

where $\{\sigma(\Delta t)\}^2 = \sigma(0)^2 + \imath\Delta t\hbar/M_{QM}$ and M_{QM} is the mass of the quantum mechanical particle. Equation 2 utilizes the well-known analytical expression for the free-propagation of a Gaussian function with spread $\sigma(0)$, [82] as well as the fact that Hermite functions, $\{H_n(x)\}$, are generated from Gaussian functions.[50] The result is a banded-Toeplitz matrix representation for the quantum propagator.[78−81] [The structure of a Toeplitz matrix is such that the (i, j)th element depends only on $|i - j|$, allowing for an efficient computational scheme that only requires the first (banded) row of the matrix to be stored. This is exploited in the "DAF" free-propagator[67] to reduce computational cost.] It is routine to carry out QWAIMD using Hermite functions of the order of $M_{DAF} = 20-30$. When using a larger number of Hermite functions, numerical stability becomes an issue, but this is surmounted through a minor modification of the recursion relation as outlined in ref 46.

II.A. Time-Dependent Deterministic Sampling (TDDS) Based QWAIMD. The evolution of the classical nuclei involves the wavepacket-averaged Hellmann−Feynman forces obtained from electronic structure calculations on the discrete wavepacket grid. To minimize the number of such calculations while maintaining accuracy, a time-dependent deterministic sampling (TDDS) function was introduced in refs 51 and 52. The mathematical form of the TDDS function is defined to be a function of the quantum nuclear degrees of freedon, R_{QM}, as follows: The TDDS function is chosen to be directly proportional to the wavepacket probability density, $\rho(R_{QM})$, and gradient of the potential, $V'(R_{QM})$, while being inversely proportional to the potential, $V(R_{QM})$, as noted in eq 3. Large values of the TDDS function represent areas where sampling should occur. The construction of TDDS has physical justifications that ensure that both classical and quantum (tunneling) regions of the dynamics are equally sampled. This gives a sampling function of the form:

$$\omega_0(R_{QM}) \propto \frac{[\tilde{\rho}(R_{QM}) + 1/I_\chi] \times [\tilde{V}'(R_{QM}) + 1/I_{V'}]}{\tilde{V}(R_{QM}) + 1/I_V} \quad (3)$$

where $\tilde{\rho}$, $\tilde{V}'$, and $\tilde{V}$ are shifted, normalized, and maintained positive semidefinite[51,52] according to:

$$\tilde{V}(R_{QM}) \propto \frac{V(R_{QM}) - V_{min}}{V_{max} - V_{min}} \quad (4)$$

and similarly for $\tilde{\rho}(R_{QM})$ and $\tilde{V}'(R_{QM})$. The quantities V_{max} and V_{min} are the maximum and minimum values for the potential, respectively, and the overall sampling function, $\omega_0(R_{QM})$, is L^1-normalized according to

$$||\omega_0(R_{QM})||_1 = \int |\omega_0(R_{QM})|\, dR_{QM} = 1 \quad (5)$$

In ref,[51] a detailed algorithm for implementation of TDDS is provided. In addition, the stability of this algorithm is also analyzed. The choice of parameters, $I_\chi = 1$, $I_{V'} = 3$, and $I_V = 1$, retains significant distribution in both the classically allowed (minimum energy regions) and classically forbidden (classical turning point) regions of the potential and leads to a large reduction in computational cost, with little perceivable loss in accuracy. The rationale behind the choice of these parameters can be qualitatively noted from the following arguments with details in ref 51. The functions $\tilde{\rho}$, $\tilde{V}'$, and $\tilde{V}$ are shifted and normalized[51] (see eq 4), and hence, (a) minimum energy regions of the potential surface are characterized by low potential energy, low gradient, and relatively high wavepacket distribution, while (b) quantum tunneling (or classical turning point) regions of the potential are approximately characterized by moderately large values of the potential, high gradients, and smaller wavepacket values. When one enforces the condition that the TDDS function must be approximately equal in these two situations for minimal bias between the classically allowed and classically forbidden regions, it is found that $I_\chi = 1$, $I_{V'} = 3$, and $I_V = 1$ provides the lowest order solution satisfying these considerations.[51] (Higher order solutions further increase the sampling in the classically forbidden regions.) In addition to these formal considerations, the parameters have been numerically tested in ref 51 for a set of 70 analytical and numerical potentials, and the results are found to be consistent with the above physical arguments. In ref 52, the TDDS implementation of QWAIMD has been benchmarked for accuracy in computing vibrational properties in hydrogen-bonded clusters. Specifically, the $ClHCl^-$ system was treated since it provides significant challenges for accurate modeling of electron−nuclear coupling.[52,83−85] In ref 52, the TDDS implementation of QWAIMD was found to accurately reproduce the experimental spectrum at limited computational cost. The analysis of trajectories is facilitated through the introduction of a novel velocity-flux correlation function.[52]

The computational implementation of TDDS[52] is achieved as follows: For quantum dyanmics beyond one dimension, the TDDS function on the full grid is evaluated at every instant in time to determine the grid points where the potential and gradients are to be obtained for the next time step. For this purpose, the TDDS function is written as a linear combination of Haar wavelets:[52]

$$\omega(x) \propto \frac{[\tilde{\rho} + 1/I_\chi] \times \left[\tilde{V}' + 1/I_{V'}\right]}{\tilde{V} + 1/I_V} = \sum_{i=0}^{N_{GEN}} \underbrace{\sum_{j_1=0}^{a^i-1} \cdots \sum_{j_{N_{Dim}}=0}^{a^i-1}}_{} c_{i,\{j\}} \left\{ \prod_{k=1}^{N_{Dim}} \mathscr{H}\left(a^i x^k - \frac{j_k N_Q}{a^i}\right) \right\} \quad (6)$$

where the Haar scaling function, $\mathscr{H}(x)$ is a square function equal to 1 for $0 \le x \le 1$ and zero otherwise. The quantity N_{GEN} is the number of wavelet generations, and the underbrace below the summations is meant to indicate that there are N_{Dim} summations, $[j_1, j_2, ..., j_{N_{\mathrm{Dim}}}]$. $c_{i,\{j\}}$ implies that the coefficients depend on i and the entire set of j indices. The Haar wavelets, $\{\mathscr{H}(a^i x - j_k N_Q/a^i)\}$, comprise a hierarchy of translated and dilated forms of $\mathscr{H}(x)$. Only the Haar scaling function is used since the Haar wavelet function is the orthogonal complement of the Haar scaling function and is not positive semidefinite, which is one of the requirements on ω. The quantity x^k, in eq 6, is the kth component of the N_{Dim} dimensional vector, and a is chosen to be 2 or 3. That is, we employ 2- and 3-scale functions in our scheme. Once the subset of grid points for "on-the-fly" potential energy determination is computed using the TDDS function, the value of the potential at the remaining points is obtained through Hermite curve interpolation.[86] The forces on classical atoms are subsequently determined through a low-pass filtered Lagrange interpolation technique introduced in ref 52. Time-dependent deterministic sampling has played a pivotal role in converting QWAIMD into an efficient computational tool through reduction of computational costs by about *3 to 4 orders of magnitude*.[52] (See Figure 1.)

It has also been numerically shown[51] that the TDDS function is inversely proportional to the Wentzel−Krames−Brillouin (WKB) length scale:

$$\frac{p}{\hbar} \equiv \lambda^{-1} \gg \left(\frac{1}{E - V(x)}\right)\frac{\partial V}{\partial x} \quad (7)$$

Thus, the TDDS function provides a larger sample of data points in the rapidly varying limit of the potential. Furthermore, it has been numerically shown[51,52] that the TDDS function is directly proportional to the Bohmian quantum potential.[87−99]

In addition, as noted in the Introduction, QWAIMD has been adopted to study hydrogen tunneling in enzyme active sites,[55,65] and QM/MM generalizations to the TDDS implementation of QWAIMD have also been completed.[53] In ref 100, the quantum dynamics tools from QWAIMD were used to compute the qualitative accuracy involved in classical *ab initio* molecular dynamics calculations of vibrational spectra in hydrogen bonded systems.

II.B. Further Computational Enhancements through Diabatic Extensions to QWAIMD. To further enhance the computational scaling of QWAIMD, in ref 54, we introduced a diabatic generalization. Essentially, multiple single particle electronic density matrices are simultaneously propagated through an extended Lagrangian scheme. Following this, the Slater determinantal wave functions associated with the density matrices are used to construct a nonorthogonal CI problem, which is computed on-the-fly to obtain the instantaneous adiabatic states. Computational efficiency arises through the diabatic approximation for the multiple density matrices: this essentially necessitates a limited dependence of the quantum nuclear degrees of freedom on the individual electronic density matrix states. Once this condition is enforced, it is found that two-electron integrals can be reused over the entire grid, which reduces the computational complexity in determining the potential surface enormously.

As will be discussed in the next few sections, the proposed methodological extensions using Shannon's entropy condition have multiple effects on the QWAIMD algorithm:

- An improved TDDS function is first derived and tested in section IV.A. This has direct impact on the TDDS implementation of QWAIMD.
- The TDDS functions obtained from Shannon's entropy measure are used to locate significant regions on a potential energy surface. Grid-based electronic structure basis functions are then placed on these important regions, as discussed in section IV.B. This feature leads to two further improvements in the QWAIMD methodology:
  - The introduction of the grid based electronic structure basis functions strengthens and influences the diabatic approximation discussed above and, in further detail, in ref 54.
  - The grid-based electronic structure basis functions also reduce the computational cost of each electronic structure calculation, and this in turn has an effect on the TDDS implementation of QWAIMD.

III. Time-Dependent Deterministic Sampling through Shannon Entropy Measure

As noted above, the physical justification for the form of TDDS is based on specific dynamical parameters (wavepacket probability density, potential, gradients), and in this section, we introduce additional sampling functions utilizing the concept of Shannon entropy. With reference to the TDDS-based implementation of QWAIMD, one particularly troublesome feature of TDDS is that sampling points can sometimes be placed in physically uninteresting regions of the potential during the dynamics simulation, in particular when both the potential energy and the gradient of the potential are high. These regions represent areas that are classically forbidden and also fail to demonstrate quantum behavior. While the TDDS function still performs remarkably well in improving efficiency with a negligible loss in accuracy,[51,52]the question we address in section IV.A is whether further improvements can be achieved. As will be shown in section IV.A, the new sampling functions introduced in this section provide a compressed set of sampling points and hence yield a more efficient procedure for "on-the-fly" dynamics. Furthermore, these functions also allow us to determine the positions of grid-based electronic bases in section IV.B for enhanced accuracy through diabatic extensions to QWAIMD.[54]

The general form of the total Shannon entropy of a system is given by[68]

$$-k\int_{-\infty}^{\infty} \mathrm{d}x\ \rho\log(\rho) \equiv -k\int_{-\infty}^{\infty} \mathrm{d}x\ S[\rho(x)] \tag{8}$$

where ρ is the probability density of the system, k is a proportionality constant, and we have defined the quantity in the integrand on the left as $S[\rho(x)] \cong -\rho\ \log(\rho)$. Later in this publication, we refer to $S[\rho(x)]$ as the "local Shannon entropy". This is because, while the full sum in eq 8 is the Shannon entropy,[68] $S[\rho(x)]$ implicitly depends on the local variable x. Shannon entropy has been used as a general mathematical tool to describe the information content in a system, provided there exists some probability distribution associated for the possible states of the system. A typical example of how this generic measure can apply to a physical situation is when Shannon entropy reduces to the notion of thermodynamic entropy for an ensemble of classical particles.[68] Given an ensemble of possible discrete microstates for a system, summation over all of these possible microstates gives the familiar thermodynamic entropy for the system, where k is now Boltzmann's constant. Another example deals with the use of Shannon entropy in DNA and protein structure determination, and the associated definition of complexity in biological systems.[70,71] In this case, a certain site in a DNA sequence, or an amino acid sequence, is defined to have an entropy that reflects the probability of finding different DNA bases (or individual amino acids for proteins) at that particular site. Each site, thus, has an entropy that contributes to the complexity of the organism. Entropy here is, of course, an information entropy and not a thermodynamic entropy since it pertains to the propensity of the appearance of amino acid residues (or DNA bases) at a chosen point in the sequence. In quantum mechanics, Shannon entropy is related to von Neumann entropy, when eq 8 is rewritten as

$$-kTr[\Gamma\ \log(\Gamma)] = kTr[S_{vN}[\Gamma]] \tag{9}$$

where $\Gamma \equiv |\psi\rangle\langle\psi|$ is now the density matrix associated with the system. Furthermore, in a fashion analogous to that in eq 8, we can define here a "local von Neumann entropy, " $S_{vN}[\Gamma]$. Along similar lines, semiclassical forms of entropy have been defined[101] where coherent states[102] have been employed for the probability function, ρ, in eq 8.

Influenced by this early work, here we utilize the local value of Shannon information entropy defined in eq 8: $S[\rho(x)] \equiv -\rho\log(\rho)$, where ρ is chosen as the time-dependent wavepacket density in our QWAIMD simulations, to construct suitable sampling functions of the form

$$\omega_0(R_{\mathrm{QM}}) = \frac{(\tilde{\rho}(R_{\mathrm{QM}}) + 1/I_\chi)(\tilde{V}'(R_{\mathrm{QM}}) + 1/I_{V'})}{(\tilde{V}(R_{\mathrm{QM}}) + 1/I_V)} \tag{10}$$

$$\omega_1(R_{\mathrm{QM}}) = \frac{(\tilde{S}[\rho(R_{\mathrm{QM}})] + 1/I_S)}{(\tilde{V}(R_{\mathrm{QM}}) + 1/I_V)} \tag{11}$$

$$\omega_2(R_{\mathrm{QM}}) = (\tilde{S}[\rho(R_{\mathrm{QM}})] + 1/I_S) \tag{12}$$

where ω_0 is the original TDDS function and ω_1 is a composite function that utilizes the Shannon entropy as well as the potential energy. In all cases, $\tilde{S}$ and $\tilde{V}$ are shifted and normalized as per eq 4, and the sampling functions are scaled such that the respective values are bounded by unity (see eq 5). In addition, while the parameters $I_\chi = 1$, $I_{V'} = 3$, and $I_V = 1$ define ω_0, those for ω_1 and ω_2 are chosen as $I_S = I_V = 1$ in this study. This allows consistent treatment of the wavepacket and the local Shannon entropy in the sampling functions. It must also be noted that the quantum mechanical Shannon entropy defined here for use in eqs 11 and 12 is a special form of the semi-classical entropy defined in ref 101. In that case, coherent states[102] were used to define the probability function instead of the time-dependent wavepacket density that is used here.

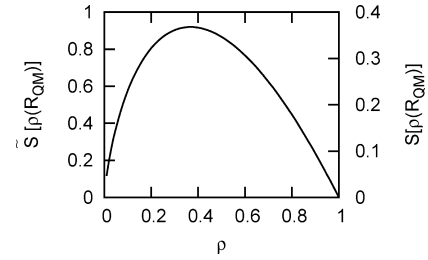


Figure 2. Behavior of $\tilde{S}$ and S as a function of ρ.

Before we proceed into a numerical analysis of these sampling functions, it is useful to inspect limiting cases for both $S[\rho(R_{\mathrm{QM}})]$ and $\tilde{S}[\rho(R_{\mathrm{QM}})]$. An illustration of the behavior of $\tilde{S}$ and S as a function of ρ is provided in Figure 2. When the probability associated with the wavepacket is high, there is information indicating the presence of the "particle" in the given region of configurational space and, hence, the entropy at that point, $S[\rho(x)] \equiv -\rho\log(\rho) \approx 0$. In a similar fashion, we note that when the probability is low there is information indicating the absence of the particle in the given region of configurational space, and consequently the local entropy, $S[\rho(x)]$, and its scaled form, $\tilde{S}[\rho(x)]$, are both expected to be small. Intermediate values of the probability function yield greater uncertainty in regard to the presence of the particle. This uncertainty may be qualitatively related to Shannon entropy, and as a result, the local entropy, $S[\rho(x)]$, and its scaled form, $\tilde{S}[\rho(x)]$ are both higher for intermediate values of ρ.

This naturally creates the situation where a sampling function constructed from $\tilde{S}[\rho(x)]$ alone, that is eq 12, has the effect of producing a higher distribution of sampling points in regions where the wavepacket amplitude is intermediate. While this may be desirable to represent tunneling regions, the regions that are classically populated may have larger ρ values that are not expected to be populated well enough when $\tilde{S}[\rho(x)]$ alone is used in a sampling function. Consequently, eq 11 has been introduced as a hybrid sampling function that includes the potential to also represent the classically allowed regions. Indeed, as we will see in a later section, it is the sampling function in eq 11 that shows the best performance of the three considered above.

In the next section, we gauge the utility of these sampling functions in probing important regions of the potential

Table 1. Energy Conservation Data from a One-Dimensional Dynamical Treatment of the Shared Proton in $[Cl-H-Cl]^{-}$[a]

level of theory	TDDS	N_Q[b]	N_E[c]	N_Q/N_E[d]	temp (K)[e]	time (ps)	ΔE (kcal/mol)
HF/6-31G(d,p)	–[f]	101	101	1	325.26	1.0	0.03
HF/6-31G(d,p)	ω_0	101	11	9.18	325.26	1.9	0.02
HF/6-31G(d,p)	ω_1	101	11	9.18	318.87	1.3	0.02
HF/6-31G(d,p)	ω_2	101	11	9.18	319.25	1.3	0.02
HF/6-31G(d,p)	ω_0	101	9	11.22	340.85	2.5	0.13
HF/6-31G(d,p)	ω_1	101	9	11.22	320.01	3.2	0.13
HF/6-31G(d,p)	ω_2	101	9	11.22	337.13	3.3	0.12
HF/6-31G(d,p)	ω_0	101	7	14.42	368.37	2.6	0.23
HF/6-31G(d,p)	ω_1	101	7	14.42	370.04	2.7	0.30
HF/6-31G(d,p)	ω_2	101	7	14.42	341.14	1.5	0.11
B3LYP/6-31+G(d,p)	–	101	101	1	258.45	1.1	0.01
B3LYP/6-31+G(d,p)	ω_0	101	11	9.18	257.63	1.7	0.06
B3LYP/6-31+G(d,p)	ω_1	101	11	9.18	261.08	0.4	0.00
B3LYP/6-31+G(d,p)	ω_2	101	11	9.18	261.30	0.4	0.00
B3LYP/6-31+G(d,p)	ω_0	101	9	11.22	261.94	2.6	0.02
B3LYP/6-31+G(d,p)	ω_1	101	9	11.22	258.59	1.8	0.03
B3LYP/6-31+G(d,p)	ω_2	101	9	11.22	260.45	1.8	0.02
B3LYP/6-31+G(d,p)	ω_1	101	7	14.42	251.23	2.4	0.10
B3LYP/6-31+G(d,p)	ω_2	101	7	14.42	256.69	4.1	0.05

[a] For all calculations, the quantum dynamical time step $\Delta t_{QM} = 0.05$ fs and the classical time-step $\Delta t_{Cl} = 0.25$ fs. [b] The total number of grid points. [c] The number of points on the grid where electronic structure calculations are performed. This set of points is obtained from TDDS and is adaptive (that is, time-dependent). [d] Represents the computational gain from TDDS. [e] Calculated from classical nuclear velocities and wavepacket kinetic energy. [f] No sampling. Electronic structure calculations performed on the full grid.

surface, both for quantum dynamics and electronic structure as stated above.

IV. Numerical Tests on Accuracy and Efficiency of the Shannon Information Entropy-Based Sampling Techniques

IV.A. Improvements to "On-the-Fly" TDDS-Based Quantum Dynamics. To evaluate the Shannon information entropy-based functions as effective TDDS functions, we compare the performance of ω_1 and ω_2 to that of ω_0. QWAIMD simulations using these sampling functions were conducted on the bihalide cluster, $[Cl-H-Cl]^{-}$. The choice of system is based on the known challenges this system presents to accurately compute electron–nuclear coupling.[52,83–85] This model system has been the subject of substantial experimental and theoretical study[83,103–109] and has been used for previous TDDS studies under QWAIMD.[51,52] The bihalide system contains a shared proton undergoing exchange between donor and acceptor atoms, allowing the possibility of proton modes to couple with the other atoms in the system. Here, we utilize this system to evaluate the effectiveness of the three sampling functions presented in the previous section. For a detailed description of the vibrational properties of this system, obtained using QWAIMD, please see ref 52. The shared proton is treated using quantum dynamics, whereas all other atoms are treated with Born–Oppenheimer molecular dynamics (BOMD), as allowed within QWAIMD. The electronic structure calculations are treated with both Hartree–Fock and DFT methods. For all Hartree–Fock simulations, 6-31G(d,p) is used as the basis set, and for DFT simulations the B3LYP functional is used alongside the 6-31+G(d,p) basis set. All QWAIMD computations in this publication are performed using a development version of the Gaussian series of electronic structure codes.[110]

Table 1 provides a summary of energy conservation data when all three sampling functions are used with QWAIMD. While using a Hartree–Fock treatment of the electronic structure, all of the sampling functions appear to perform well with 11 sampling points per dimension leading to an order of magnitude compression of the quantum grid. But it is also noted that using seven sampling points per dimension leads to acceptable results.

Figures 3 and 4 qualitatively demonstrate the effectiveness of the Shannon entropy based sampling functions. In Figure 3, we present the evolution of all of the sampling functions computed from dynamics data calculated using ω_0. It is already clear that there are differences in the way ω_0 samples the edges of the grid as compared to ω_1 and ω_2. For example, note that the edges of the grid are much darker for the case of ω_1 as compared to ω_0. This important difference is further highlighted in Figure 4, where again it is noted that ω_0 shows a higher density at the ends of the grid as compared to ω_1 and ω_2. Furthermore, the center of the grid is sampled to a slightly greater extent by ω_1, although all sampling functions sample this region suitably. These results are consistent with the discussions at the end of section III, where we expected ω_1 to provide a greater sample in the classically allowed regions as compared to ω_2. However, the fact that both ω_1 and ω_2 provide a reduced sampling at the grid edges arises due to there being no functional dependence on V' in the cases of ω_1 and ω_2.

To further quantify the differences between the sampling functions, the overlapping regions between the sampling functions are calculated at each step using

$$\omega_i'(x_j;t) = \omega_i(x_j;t)\,\omega_0(x_j;t) \tag{13}$$

where x_j is a particular grid point, and $i = 1$ and 2; i.e., ω_i above represents one of the Shannon entropy based functions. The evolution of eq 13, provided in Figure 5, shows the common and uncommon regions sampled as the functions

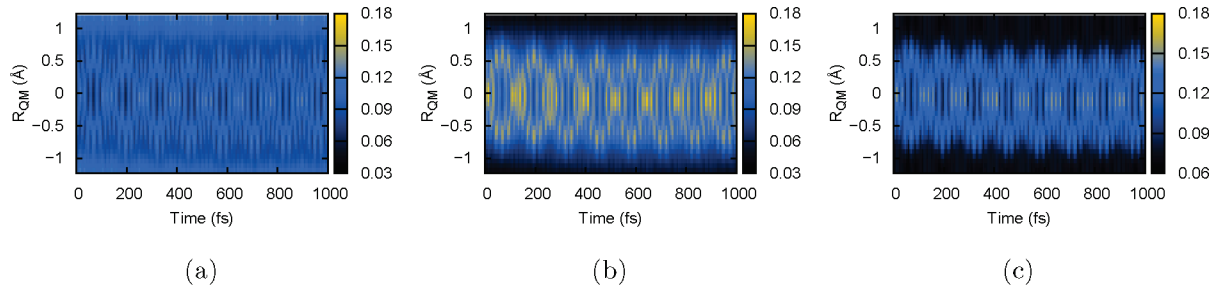


Figure 3. A comparison of the sampling functions, ω_0 (a), ω_1 (b), and ω_2 (c). The figures depict the evolution of the sampling functions during a single reference dynamics trajectory. The intensity of ω_0 is relatively high at the edges of the grid as compared to both ω_1 and ω_2. Similarly, the intensity of ω_1 is higher in the important regions as compared to both ω_0 and ω_2. Note that this is not a comparison of actual dynamical data. ω_1 and ω_2 were reconstructed using dynamics data performed with ω_0.

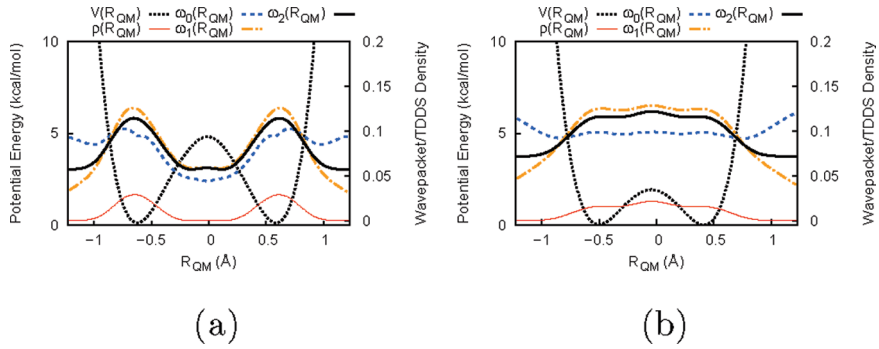


Figure 4. (a) A representative time slice of the sampling functions, ρ, and the potential from a one-dimensional Hartree−Fock simulation. (b) The time-averaged behavior. Again, as already seen in Figure 3, the fact that ω_0 overestimates the significance of the edges of the grid is clearly noted. Furthermore, ω_1 has higher intensities in the important regions, consistent with Figure 3.

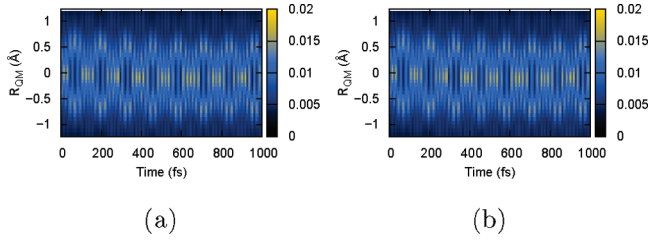


Figure 5. Evolution of eq 13 for ω_1 (a) and ω_2 (b). The common regions between ω_0 and the new TDDS functions are represented by regions of high intensity. These are located in the central regions of the grid. The lower intensity at the edges demonstrates the crucial difference between the two sets of sampling functions.

change form due to the dynamics. Regions of high density of the overlap measure correspond to regions of high commonality, while low density shows areas of divergent behavior between the two sampling functions. As seen in the time-averaged behavior of the sampling functions in Figure 4b, the common regions are contained in the center of the grid. However, now the time-dependency of this relationship is revealed, and major oscillations in these functions are preserved. Furthermore, the fact that edges of the grid are lighter in Figure 5b as compared to Figure 5a indicates a greater sampling of the grid edges for the case of ω_2 as compared to ω_1.

In the TDDS algorithm, once the sampling function is constructed, a Haar wavelet fit of this function is discretized to obtain points in configurational space for electronic structure calculations. (The detailed algorithm is presented in refs 52 and 51.) The discrete, time-dependent set of points obtained from such an algorithm when using the functions, ω_0, ω_1, and ω_2, are shown in Figures 6 and 7. Consistent with the previous discussion, ω_1 provides the most compressed representation of the grid populating only the important regions. It is followed very closely by ω_2, and ω_0 provides a greater sampling of points at the edges of the grid. In addition, the fundamental oscillations near the center of the grid are captured by all three functions, but these oscillations are more intense for the Shannon sampling-based functions. In section IV.B, this property is used to construct a set of potential adapted, grid-based electronic structure basis functions. That is, in section IV.B, electronic structure basis functions are to be placed along the grid lines seen in Figures 6 and 7 for potential energy calculations. Such a basis set is found to be accurate and efficient and is used in ref 54 to further enhance the computational efficiency of QWAIMD.

Having examined the differences between the sampling functions, it is important to see how these directly affect observables in the dynamics. Thus, we conclude this section with an analysis of the vibrational effects on the

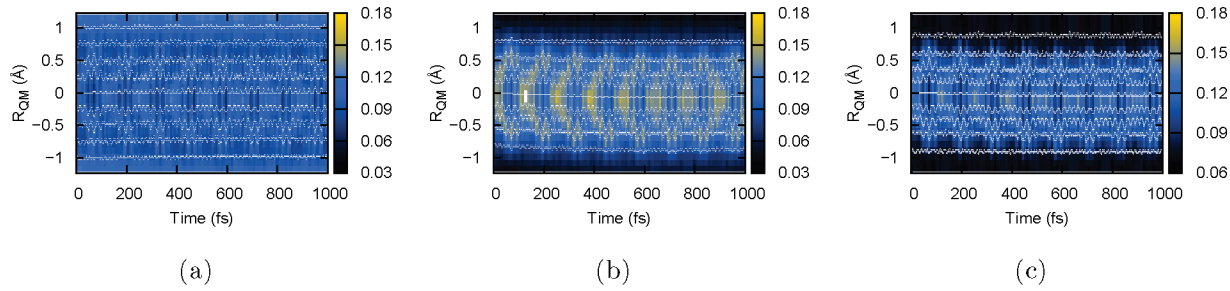


Figure 6. Time-evolution of sampling points (white lines), compared to sampling function density (blue and yellow density map), for $N_E = 11$. ω_0 is shown in a, ω_1 in b, and ω_2 in c.

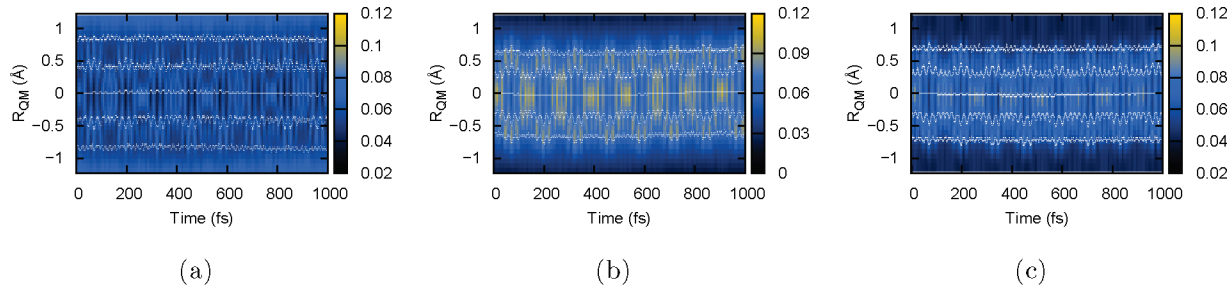


Figure 7. Time-evolution of sampling points (white lines), compared to sampling function density (blue and yellow density map), for $N_E = 7$. ω_0 is shown in a, ω_1 in b, and ω_2 in c.

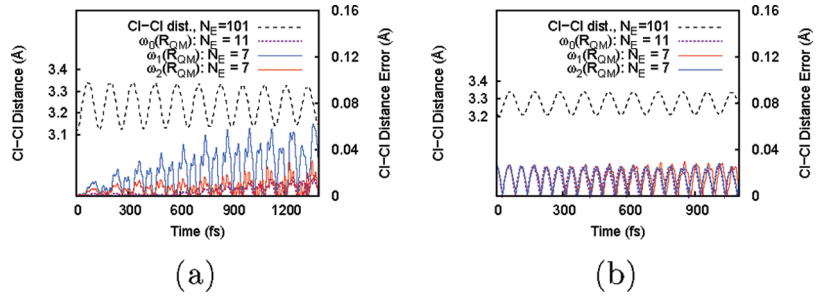


Figure 8. Error in the Cl−Cl distance. (a) Hartree−Fock simulation. (b) DFT calculation (B3LYP). The dotted black lines in both parts display the evolution of the Cl−Cl distance (left axis). The error in the Cl−Cl distance is shown on the right axis and depicted using the red and blue lines. The Cl−Cl oscillations for Hartree−Fock have a larger amplitude due to higher temperatures for the associated simulations. (Please see Table 1.)

classical atoms. The root mean squared error in the Cl−Cl distance is shown in Figure 8. The error was referenced to a QWAIMD simulation in which no interpolation of the potential and gradients was used. In all cases, the Shannon entropy based sampling functions are able to reproduce the oscillations with fewer sampling points. This, of course, is a result of the more compact nature of these sampling functions. The oscillation frequencies are in agreement with previous calculations,[52] but the result in Figure 8 indicates a reduced computational cost when using ω_1 and ω_2.

IV.B. Locating Regions for Potential-Adapted, Grid-Based Electronic Structure Basis Functions Using the Shannon Entropy Based Sampling Functions. In this section, we utilize the TDDS functions to obtain a grid-based description of electronic structure. This study is particularly relevant for hydrogen-bonded systems, and we show here that accurate potential energy surfaces can be obtained over a wide range of energies and nuclear geometries when grid-based Gaussian basis functions, directed using TDDS, are utilized. Essentially, the question we pose is, if Gaussian basis functions of the kind

$$\chi^{\mathbf{R_F}}_{l,m,n}(\mathbf{r}) = (x - R_x)^l (y - R_y)^m (z - R_z)^n \exp[-\alpha(\mathbf{r} - \mathbf{R_F})^2] \quad (14)$$

were directed such that the basis functions centers, $\mathbf{R_F}[\equiv(R_x, R_y, R_z)]$, were chosen to be functions of multiple classical nuclear variables according to $\mathbf{R_F} = f(\{\mathbf{R_C}\})$ and the centers are determined using the sampling functions, can this improve efficiency while retaining the accuracy of electronic structure calculations? In eq 14, the quantities l, m, and n are the usual orbital angular momentum indices of the basis function. The result of this discussion is a generalization of bond-centered basis functions[111−114] traditionally used in quantum chemistry where the positions of these Gaussian basis functions are determined using the TDDS functions discussed in section III. Furthermore, these

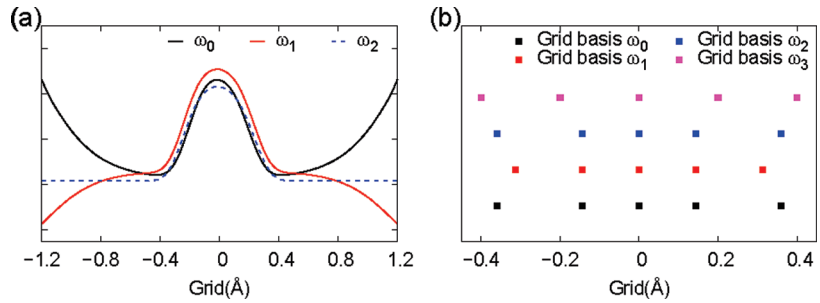


Figure 9. (a) The TDDS functions and (b) the associated origins for the grid-based electronic basis functions obtained from the TDDS functions. The system under study is $[ClHCl]^{-1}$, and the horizontal axis for both figures represents $(R_{H-Cl_1} - R_{H-Cl_2})/2$. The definitions for R_{H-Cl_1} and R_{H-Cl_2} are provided in Figure 10. Note that b shows a smaller spatial region since this is the predominant area for ω_1. Note also that the functions ω_0 and ω_2 place a greater weight at the edges, which is consistent with our earlier discussion in section IV.A.

grid-based electronic functions are used in ref 54 to further improve the efficiency of QWAIMD.

For the case of hydrogen-bonded systems, we specialize our definition of $\mathbf{R_F} \equiv f(\{\mathbf{R_C}\})$ to a function of the donor and acceptor coordinates:

$$\mathbf{R}_\mathbf{F}^i \equiv \sum_j c_{ji}\mathbf{R}_\mathbf{C}^j + \vec{\mathbf{v}}_i = a_i\mathbf{R_A} + d_i\mathbf{R_D} + \vec{\mathbf{v}}_i \qquad (15)$$

where $\mathbf{R_A}$ and $\mathbf{R_D}$ are coordinate vectors of the donor and acceptor atoms for a hydrogen-bonded system and $\vec{\mathbf{v}}_i$ is a uniform shift that can be used to create a three-dimensional grid of electronic basis functions. It is further important to note that the basis functions introduced in eq 15 are functions of classical nuclear coordinates. Hence, in a fashion similar to atom-centered basis functions, the centers of these functions also transform according to the classical nuclear positions. Furthermore, these grid-based functions are spread uniformly in space. But these functions differ from plane-waves[115] through the $\{\mathbf{R_F}\}$ dependence of the Fourier transforms.

To choose the variables $\{a_i,\ d_i,\ \vec{\mathbf{v}}_i\}$, we utilize the sampling functions discussed earlier. Our test case involves three well-studied hydrogen-bonded ion clusters:[83,103,107,108,116−124] the bihalide cluster $[ClHCl]^{-1}$, the hydroxide water cluster $[OH{-}H_2O]^{-1}$, and the Zundel cation $[H_2O{-}H{-}H_2O]^{+}$. Our goal is to find the optimum number and associated positions of the grid-based basis functions in the bonding region of the transferring hydrogen. The stability and vibrational properties of the clusters discussed here are sensitive to the potential surface along the donor−acceptor axis. Thus, potential energy surfaces constructed on a one-dimensional grid were compared between grid-based basis and atom-centered basis set aug-cc-pvtz.[125−127] To quantify the errors, we define

$$\Delta V(\varepsilon_1, \varepsilon_2) = \sqrt{\frac{\sum_i [V_1(\mathbf{R_C}, \mathbf{R}_\mathbf{QM}^\mathbf{i}) - V_2(\mathbf{R_C}, \mathbf{R}_\mathbf{QM}^\mathbf{i})]^2 \prod_{\varepsilon_1,\varepsilon_2} (V(\mathbf{R}_\mathbf{QM}^\mathbf{i}))}{\sum_i \prod_{\varepsilon_1,\varepsilon_2} (V(\mathbf{R}_\mathbf{QM}^\mathbf{i}))}} \qquad (16)$$

where the boxcar function is defined as linear combination of Heaviside functions: $\Pi_{\varepsilon_1,\varepsilon_2}(V) = H(V - \varepsilon_1) - H(V - \varepsilon_2)$.

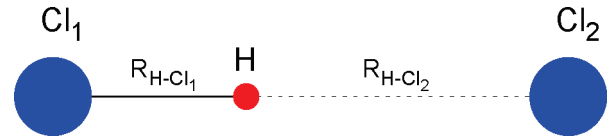


Figure 10. The parameters R_{H-Cl_1} and R_{H-Cl_2} are defined here and used in Figure 9.

Equation 16 allows us to inspect the accuracy in the potential surface in a tiered fashion by focusing on specific energy domains. We have utilized the three functions presented in eqs 10, 11, and 12 and compared the associated behavior with a uniform distribution function: $\omega_3 = 1$.

In Figure 9, distributions of the potential-adapted, grid-based basis using various TDDS schemes are presented. Compared to the uniform sampling function, ω_3, the Shannon entropy based TDDS function reduces the population of electronic basis functions close to the edge of the grid. The standard TDDS function, ω_0, on the contrary, places roughly equal weight at both the middle and edge of the grid. This, of course, is to be expected, since TDDS in eq 10 has been tuned such that the grid-based basis is distributed equally in both classically allowed and forbidden regions. However, bases at the edges (high gradients and large values of the potential) may not be useful during the electronic structure calculations, and hence, in practice one might expect the Shannon information based TDDS functions to be more efficient.

A detailed examination of the accuracy of various TDDS functions in obtaining good estimates for the potential surface is provided in Table 2. The error estimates utilized are those discussed in eq 14. Since a large number of grid-based basis functions are distributed in the bonding region of the hydrogen-bonded systems considered, a relatively small basis set (3-21G and STO-3G) is used at each grid point. The accuracy of the potential-adapted, grid-based basis functions is ascertained through comparison with a standard atom-centered aug-cc-pvtz basis. To perform the benchmark in a tiered fashion, we first replace the atom-centered aug-cc-pvtz on the shared proton with grid-based basis functions while retaining the aug-cc-pvtz bases on all of the other classical atoms. Following this, the aug-cc-pvtz bases on classical atoms are substituted with the

Table 2. Benchmarks for Grid-Based Electronic Structure Basis Functions

	atom centered basis		grid basis			error in PES[a]		
system	basis set	N_{basis}[b]	quantum/classical[c]	NB_{use} /N_{basis}[d]	$N_{grid\text{-}basis}$[e]	ΔV[f]	ΔV[g]	TDDS
$\{[ClHCl]^{-1}\}$[h]	aug-cc-pvtz	**123**	STO-3G/aug-cc-pvtz	113/121	21	0.20	0.68	ω_3
			STO-3G/aug-cc-pvtz	111/111	11	0.25	0.95	ω_3
			STO-3G[h] /aug-cc-pvtz	109/109	9	0.44	1.22	ω_3
			STO-3G/aug-cc-pvtz	111/111	11	0.26	0.92	ω_0
			STO-3G/aug-cc-pvtz	110/111	11	0.23	0.88	ω_1
			STO-3G/aug-cc-pvtz	**109/109**	**9**	**0.25**	**0.87**	$\boldsymbol{\omega_1}$
			3-21G/6-31+G**	66/86	42	0.23	0.90	ω_3
			3-21G/6-31+G**	61/66	22	0.22	1.06	ω_3
			STO-3G/6-31+G**	57/65	21	0.20	0.84	ω_3
			STO-3G/6-31+G**	55/55	11	0.20	1.15	ω_3
			STO-3G[h] /6-31+G**	53/53	9	0.44	1.46	ω_3
			STO-3G/6-31+G**	55/55	11	0.26	1.06	ω_0
			STO-3G/6-31+G**	54/55	11	0.27	1.10	ω_1
			STO-3G/6-31+G**	**53/53**	**9**	**0.27**	**1.11**	$\boldsymbol{\omega_1}$
$\{[OH-H_2O]^{-1}\}$[i]	aug-cc-pvtz	**161**	STO-3G/aug-cc-pvtz	147/147	9	0.34	0.84	ω_3
			STO-3G/aug-cc-pvtz	146/147	9	0.43	0.69	ω_0
			STO-3G/aug-cc-pvtz	145/147	9	0.36	0.79	ω_1
			STO-3G/aug-cc-pvtz	**145/145**	**7**	**0.23**	**0.85**	$\boldsymbol{\omega_1}$
			STO-3G/6-31+G**	55/55	9	0.44	1.55	ω_3
			STO-3G/6-31+G**	55/55	9	0.46	1.44	ω_0
			STO-3G/6-31+G**	53/55	9	0.48	2.11	ω_1
			STO-3G/6-31+G**	**53/53**	**7**	**0.51**	**2.10**	$\boldsymbol{\omega_1}$
$\{[H_2O-H-H_2O]^{+}\}$[j]	aug-cc-pvtz	**207**	STO-3G/aug-cc-pvtz	193/193	9	0.23	0.91	ω_3
			STO-3G/aug-cc-pvtz	192/193	9	0.26	0.86	ω_0
			STO-3G/aug-cc-pvtz	191/193	9	0.21	0.89	ω_1
			STO-3G/aug-cc-pvtz	**190/191**	**7**	**0.24**	**0.86**	$\boldsymbol{\omega_1}$
			STO-3G/6-31+G**	65/65	9	0.41	1.56	ω_3
			STO-3G/6-31+G**	64/65	9	0.32	1.42	ω_0
			STO-3G/6-31+G**	63/65	9	0.35	1.84	ω_1
			STO-3G/6-31+G**	**62/63**	**7**	**0.43**	**2.05**	$\boldsymbol{\omega_1}$

[a] Errors (in kcal/mol) are compared between atom-centered aug-cc-pvtz and grid-based basis results using eq 16. All PESs are obtained at the Hartree−Fock level. [b] Number of basis functions in the atom-centered basis calculations. [c] The shared proton is treated quantum mechanically, and all other atoms are classical. The column represents the grid basis used (around the quantum nucleus) and the atom-centered basis placed on each classical atom. [d] N_{basis} is the total number of basis functions, whereas NB_{use} is the number of linearly independent basis functions. [e] Number of basis functions used for grid-based basis. [f] $\varepsilon_1 = 0.0$, $\varepsilon_2 = 2.5$. Unit is kcal/mol. See eq 16. [g] $\varepsilon_1 = 2.5$, $\varepsilon_2 = 15.0$. Unit is kcal/mol. See eq 16. [h] Optimized geometry using MP2/aug-cc-pvtz. Cl−Cl distance is 3.13 Å. [i] Optimized geometry using MP2/aug-cc-pvtz. Oxygen−oxygen distance is 2.48 Å. [j] Optimized geometry using MP2/aug-cc-pvtz. Oxygen−oxygen distance is 2.39 Å.

relatively small double split valence 6-31+G** basis. All results are summarized in Table 2.

Although all TDDS schemes give accurate results in the low energy regions, as seen from the smaller values of ΔV in the column using $\varepsilon_1 = 0.0$ kcal/mol and $\varepsilon_2 = 2.5$ kcal/mol, the Shannon entropy based TDDS function, ω_1, provides higher accuracy while using fewer basis functions. (These are shown in blue in Table 2.) The reduction in the number of grid-based basis functions is especially striking in this case where the number of basis functions required is reduced to roughly half in the case of $[ClHCl]^{-1}$ and a third in the case of the larger systems. There appears to be little loss in accuracy over the entire grid. Due to the $O(N^3)$ scaling of the algorithms involved, this leads to a factor of 8 reduction in computation time for the smaller $[ClHCl]^{-1}$ system and a factor of 27 reduction in computation time for the larger systems. In ref 54, these potential-adapted grid-based electronic basis functions are utilized to facilitate an even larger reduction in computation time when employed in conjunction with new formalisms of QWAIMD.

V. Conclusions

A new set of time-dependent deterministic sampling functions based on Shannon's entropy were introduced. These functions were used to probe important regions of an electronic potential surface and to facilitate computational improvements in quantum-classical dynamics of electrons and nuclei. Computational gains are two-fold as discussed in the numerical results section: The direct implementation of Shannon entropy based TDDS functions reduces computational cost by eliminating the need for sampling points in physically uninteresting regions of the potential surface. In addition, when the Shannon entropy based TDDS functions are utilized to construct a potential-adapted grid-based electronic basis set, the accuracy of the electronic potential surface is well-preserved, while the computational cost is significantly lowered. This idea is further exploited in ref 54 to facilitate the development of a new QWAIMD formalism that reduces computational costs by several orders of magnitude.

Acknowledgment. This research is supported by the National Science Foundation, grant number CHE-0750326 to S.S.I.

References

(1) Wyatt, R. E.; Zhang, J. Z. H. *Dynamics of Molecules and Chemical Reactions*; Marcel Dekker Inc.: New York, 1996.

(2) Berne, B. J.; Ciccotti, G.; Coker, D. F. *Classical and Quantum Dynamics in Condensed Phase Simulations*; World Scientific: Singapore, 1997.

(3) Wang, I. S. Y.; Karplus, M. *J. Am. Chem. Soc.* **1973**, *95*, 8160.

(4) Leforestier, C. *J. Chem. Phys.* **1978**, *68*, 4406.

(5) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(6) Bolton, K.; Hase, W. L.; Peslherbe, G. H. World Scientific: Singapore, 1998; Chapter: Direct Dynamics of Reactive Systems, p 143.

(7) Schlegel, H. B.; Millam, J. M.; Iyengar, S. S.; Voth, G. A.; Daniels, A. D.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2001**, *114*, 9758.

(8) Schatz, G. C.; Kupperman, A. *J. Chem. Phys.* **1976**, *65*, 4642.

(9) Delos, J. B. *Rev. Mod. Phys.* **1981**, *53*, 287.

(10) Feit, M. D.; Fleck, J. A. *J. Chem. Phys.* **1982**, *78*, 301.

(11) Kosloff, R. *Annu. Rev. Phys. Chem.* **1994**, *45*, 145.

(12) Leforestier, C.; Bisseling, R. H.; Cerjan, C.; Feit, M. D.; Freisner, R.; Guldberg, A.; Hammerich, A.; Jolicard, D.; Karrlein, W.; Meyer, H. D.; Lipkin, N.; Roncero, O.; Kosloff, R. *J. Comput. Phys.* **1991**, *94*, 59.

(13) DeVries, P. *Atomic and molecular processes with short intense laser pulses*; Bandrauk, A. D., Ed.; Plenum Press: New York, 1988; Vol. 171 of NATO ASI Series B, Physics, p 481.

(14) Jang, H. W.; Light, J. C. *J. Chem. Phys.* **1995**, *102*, 3262–3268.

(15) Althorpe, S. C.; Clary, D. C. *Annu. Rev. Phys. Chem.* **2003**, *54*, 493–529.

(16) Huang, Y.; Iyengar, S. S.; Kouri, D. J.; Hoffman, D. K. *J. Chem. Phys.* **1996**, *105*, 927.

(17) Miller, W. H.; Schwartz, S. D.; Tromp, J. W. *J. Chem. Phys.* **1983**, *79*, 4889.

(18) Makri, N. *Comput. Phys. Commun.* **1991**, *63*, 389–414.

(19) Cao, J.; Voth, G. A. *J. Chem. Phys.* **1994**, *100*, 5093.

(20) Cao, J.; Voth, G. A. *J. Chem. Phys.* **1994**, *100*, 5106.

(21) Cao, J.; Voth, G. A. *J. Chem. Phys.* **1994**, *101*, 6168.

(22) Jang, S.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 2357.

(23) Jang, S.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 2371.

(24) Feit, M. D.; Fleck, J. A. *J. Chem. Phys.* **1983**, *79*, 301.

(25) Feit, M. D.; Fleck, J. A. *J. Chem. Phys.* **1984**, *80*, 2578.

(26) Kosloff, D.; Kosloff, R. *J. Comput. Phys.* **1983**, *52*, 35.

(27) Kosloff, D.; Kosloff, R. *J. Chem. Phys.* **1983**, *79*, 1823.

(28) Tal-Ezer, H.; Kosloff, R. *J. Chem. Phys.* **1984**, *81*, 3967.

(29) Hartke, B.; Kosloff, R.; Ruhman, S. *Chem. Phys. Lett.* **1986**, *158*, 223.

(30) Iyengar, S. S.; Kouri, D. J.; Hoffman, D. K. *Theor. Chem. Acc.* **2000**, *104*, 471.

(31) Lill, J. V.; Parker, G. A.; Light, J. C. *Chem. Phys. Lett.* **1982**, *89*, 483.

(32) Light, J. C.; Hamilton, I. P.; Lill, J. V. *J. Chem. Phys.* **1985**, *82*, 1400.

(33) Colbert, D. T.; Miller, W. H. *J. Chem. Phys.* **1992**, *96*, 1982–1991.

(34) Huang, Y.; Kouri, D. J.; Arnold, M.; Thomas, L.; Marchioro, I.; Hoffman, D. K. *Comput. Phys. Commun.* **1994**, *80*, 1.

(35) Deumens, E.; Diz, A.; Longo, R.; Öhrn, Y. *Rev. Mod. Phys.* **1994**, *66*, 917.

(36) Hack, M. D.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 7917–7926.

(37) Miller, W. H. *J. Phys. Chem. A* **2001**, *105*, 2942–2955.

(38) Heller, E. J. *J. Chem. Phys.* **1975**, *62*, 1544–1555.

(39) Fiete, G. A.; Heller, E. J. *Phys. Rev. A* **2003**, *68*, 022112.

(40) Hammes-Schiffer, S.; Tully, J. *J. Chem. Phys.* **1994**, *101*, 4657–4667.

(41) Martinez, T. J.; Ben-Nun, M.; Ashkenazi, G. *J. Chem. Phys.* **1996**, *104*, 2847.

(42) Micha, D. A. *J. Phys. Chem. A* **1999**, *103*, 7562–7574.

(43) Payne, M. C.; Teter, M. P.; Allan, D. C.; Arias, T. A.; Joannopoulos, J. D. *Rev. Mod. Phys.* **1992**, *64*, 1045.

(44) Marx, D.; Hutter, J. John von Neumann Institute for Computing: Julich, Germany, 2000; Chapter: Ab Initio Molecular Dynamics: Theory and Implementation, Vol. 1, pp 301−449.

(45) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514–1527.

(46) Iyengar, S. S.; Jakowski, J. *J. Chem. Phys.* **2005**, *122*, 114105.

(47) Pavese, M.; Berard, D. R.; Voth, G. A. *Chem. Phys. Lett.* **1999**, *300*, 93–98.

(48) Tuckerman, M. E.; Marx, D. *Phys. Rev. Lett.* **2001**, *86*, 4946–4949.

(49) Chen, B.; Ivanov, I.; Klein, M. L.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *91*, 215503.

(50) Iyengar, S. S. *Theor. Chem. Acc.* **2006**, *116*, 326.

(51) Jakowski, J.; Sumner, I.; Iyengar, S. S. *J. Chem. Theory Comput.* **2006**, *2*, 1203–1219.

(52) Sumner, I.; Iyengar, S. S. *J. Phys. Chem. A* **2007**, *111*, 10313–10324.

(53) Sumner, I.; Iyengar, S. S. *J. Chem. Phys.* **2008**, *129*, 054109.

(54) Li, X.; Iyengar, S. S. *J. Chem. Phys.* **2010**, *132*, 184105.

(55) Iyengar, S. S.; Sumner, I.; Jakowski, J. *J. Phys. Chem. B* **2008**, *112*, 7601.

(56) Iyengar, S. S. *Int. J. Quantum Chem.* **2009**, *109*, 3798.

(57) Tully, J. C. *Faraday Discuss.* **1998**, *110*, 407–419.

(58) Kapral, R.; Ciccotti, G. *J. Chem. Phys.* **1999**, *110*, 8919.

(59) Horenko, I.; Salzmann, C.; Schmidt, B.; Schutte, C. *J. Chem. Phys.* **2002**, *117*, 11075–11088.

(60) Donoso, A.; Zheng, Y. J.; Martens, C. C. *J. Chem. Phys.* **2003**, *119*, 5010.

(61) Brooksby, C.; Prezhdo, O. V. *Chem. Phys. Lett.* **2001**, *346*, 463–469.

(62) Prezhdo, O. V.; Brooksby, C. *Phys. Rev. Lett.* **2000**, *86*, 3215–3219.

(63) Gindensperger, E.; Meier, C.; Beswick, J. A. *J. Chem. Phys.* **2000**, *113*, 9369.

(64) Li, X.; Iyengar, S. S. http://www.indiana.edu/ssiweb/papers/sadafp.pdf. Submitted to *J. Phys. Chem. A*.

(65) Sumner, I.; Iyengar, S. S. *J. Chem. Theory Comput.* **2010**, *6*, 1698.

(66) Pacheco, A.; Iyengar, S. S. *J. Chem. Phys.* **2010**, *133*, 044105.

(67) Pacheco, A.; Iyengar, S. S. *J. Chem. Phys.* In Press. http://www.indiana.edu/ssiweb/papers/msaiwd2.pdf (accessed Dec 2010).

(68) Shannon, C. *Bell Syst. Tech. J.* **1948**, *27*, 279–423.

(69) Shannon, C. *Proc. IEEE* **1998**, *86*, 447.

(70) C, C. A.; Ofria, C.; Collier, T. C. *Proc. Natl. Acad. Sci.* **2000**, *97*, 4463.

(71) Schneider, T. D.; Stormo, G. D.; Gold, L.; Ehrenfeucht, A. *J. Mol. Biol.* **1986**, *188*, 415.

(72) Wehner, S.; Winter, A. *J. Math. Phys.* **2008**, *49*, 062105.

(73) Iyengar, S. S.; Frisch, M. J. *J. Chem. Phys.* **2004**, *121*, 5061.

(74) Iyengar, S. S.; Schlegel, H. B.; Millam, J. M.; Voth, G. A.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2001**, *115*, 10291.

(75) Trotter, M. F. *Proc. Am. Math. Soc.* **1959**, *10*, 545.

(76) Nelson, E. *J. Math. Phys.* **1964**, *5*, 332.

(77) Strang, G. *SIAM J. Numer. Anal.* **1968**, *5*, 506–516.

(78) Korevaar, J. *Am. Math. Soc. Trans.* **1959**, *91*, 53–101.

(79) Kouri, D. J.; Huang, Y.; Hoffman, D. K. *Phys. Rev. Lett.* **1995**, *75*, 49–52.

(80) Hoffman, D. K.; Nayar, N.; Sharafeddin, O. A.; Kouri, D. J. *J. Phys. Chem.* **1991**, *95*, 8299.

(81) Yu, S.; Zhao, S.; Wei, G. W. *J. Comput. Phys.* **2005**, *206*, 727–780.

(82) Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics and Path Integrals*; McGraw-Hill Book Company: New York, 1965.

(83) Swalina, C.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2005**, *109*, 10410.

(84) Gerber, R. B.; Ratner, M. A. *J. Chem. Phys.* **1988**, *70*, 97–132.

(85) Matsunaga, N.; Chaban, G. M.; Gerber, R. B. *J. Chem. Phys.* **2002**, *117*, 3541.

(86) Bartels, R. H.; Beatty, J. C.; Barsky, B. A. *An Introduction to Splines for use in computer graphics and geometric modeling*; Morgan Kaufman Publishers, Inc.: Los Altos, CA, 1987.

(87) Sakurai, J. J. *Modern Quantum Mechanics*; Addison-Wesley Publishing Company: Reading, MA, 1994.

(88) Madelung, E. *Z. Phys.* **1926**, *40*, 322–326.

(89) de Broglie, L. *An introduction to the study of wave mechanics*; E. P. Dutton and Company, Inc.: New York, 1930.

(90) Bohm, D. *Quantum Theory*; Prentice-Hall Inc.: New York, 1951.

(91) Bohm, D. *Phys. Rev.* **1952**, *85*, 166.

(92) Cushing, J. T.; Fine, A.; Goldstein, S. *Bohmian Mechanics: An appraisal*; Kluwer: Boston, 1996.

(93) Holland, P. R. *The Quantum Theory of Motion*; Cambridge, New York, 1993.

(94) Lopreore, C. L.; Wyatt, R. E. *Phys. Rev. Lett.* **1999**, *82*, 5190.

(95) Day, B. K.; Askar, A.; Rabitz, H. A. *J. Chem. Phys.* **1998**, *109*, 8770.

(96) Wyatt, R. E.; Kouri, D. J.; Hoffman, D. K. *J. Chem. Phys.* **2000**, *112*, 10730.

(97) Bittner, E. R.; Wyatt, R. E. *J. Chem. Phys.* **2000**, *113*, 8888.

(98) Wyatt, R. E.; Bittner, E. R. *Comput. Sci. Eng.* **2003**, *5*, 22–30.

(99) Iyengar, S. S.; Schlegel, H. B.; Voth, G. A. *J. Phys. Chem. A.* **2003**, *107*, 7269–7277.

(100) Li, X.; Oomens, J.; Eyler, J. R.; Moore, D. T.; Iyengar, S. S. *J. Chem. Phys.* **2010**, *132*, 244301.

(101) Wehrl, A. *Rev. Mod. Phys.* **1978**, *50*, 221.

(102) Klauder, J. R.; Skagerstam, B.-S. *Coherent States: Applications in Physics and Mathematical Physics*; World Scientific Publishing Company, Inc.: River Edge, NJ, 1985.

(103) Kawaguchi, K. *J. Chem. Phys.* **1988**, *88*, 4186–4189.

(104) Swalina, C.; Pak, M. V.; Hammes-Schiffer, S. *Chem. Phys. Lett.* **2005**, *404*, 394.

(105) McCoy, A. B. *J. Chem. Phys.* **1995**, *103*, 986.

(106) Botschwina, P.; Sebal, P.; Burmeister, R. *J. Chem. Phys.* **1988**, *88*, 5246.

(107) Metz, R. B.; Kitsopoulos, T.; Weaver, A.; Neumark, D. *J. Chem. Phys.* **1988**, *88*, 1463.

(108) Del Popolo, M. G.; Kohanoff, J.; Lynden-Bell, R. M. *J. Phys. Chem. B* **2006**, *110*, 8798.

(109) Mo, O.; Yanez, M.; Del Bene, J. E.; Alkorta, L.; Elguero, J. *ChemPhysChem* **2005**, *6*, 1411.

(110) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian Development Version*, revision B.01; Gaussian, Inc.: Pittsburgh, PA, 2010.

(111) Rothenberg, S.; Schaefer, H. F., III. *J. Chem. Phys.* **1971**, *54*, 2764.

(112) Tao, F.-M.; Pan, Y.-K. *J. Chem. Phys.* **1992**, *97* (7), 4989–4995.

(113) Tao, F.-M. *J. Chem. Phys.* **1993**, *98* (3), 2481–2483.

(114) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103* (17), 7374–7391.

(115) Füsti-Molnar, L.; Pulay, P. *J. Chem. Phys.* **2002**, *116*, 7795.

(116) Kreuer, K. D.; Fuchs, A.; Ise, M.; Spaeth, M.; Maier, J. *Electrochim. Acta* **1998**, *43*, 1281–1288.

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

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

(119) Tuckerman, M. E.; Marx, D.; Parrinello, M. *Nature* **2002**, *417*, 925–929.

(120) Marx, D.; Tuckerman, M. E.; Hutter, J.; Parrinello, M. *Nature* **1999**, *397*, 601–604.

(121) Asthagiri, D.; Pratt, L. R.; Kress, J. D.; Gomez, M. A. *Proc. Natl. Acad. Sci.* **2004**, *101*, 7229–7233.

(122) Shin, J.-W.; Hammer, N. I.; Diken, E. G.; Johnson, M. A.; Walters, R. S.; Jaeger, T. D.; Duncan, M. A.; Christie, R. A.; Jordan, K. D. *Science* **2004**, *304*, 1137–1140.

(123) Diken, E. G.; Headrick, J. M.; Roscioli, J. R.; Bopp, J. C.; Johnson, M. A.; McCoy, A. B.; Huang, X.; Carter, S.; Bowman, J. M. *J. Phys. Chem. A* **2005**, *109*, 571–575.

(124) Hammer, N. I.; Diken, E. G.; Roscioli, J. R.; Johnson, M. A.; Myshakin, E. M.; Jordan, K. D.; McCoy, A. B.; Huang, X.; Bowman, J. M.; Carter, S. *J. Chem. Phys.* **2005**, *122* (24), 244301.

(125) Dunning Jr, T. H. *J. Chem. Phys.* **1989**, *90* (2), 1007–1023.

(126) Mourik, T. V.; Wilson, A. K.; Dunning, T. H., Jr. *Mol. Phys.* **1999**, *96* (4), 529–547.

(127) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98* (2), 1358–1371.

CT1005856

JCTC Journal of Chemical Theory and Computation

Reaction Ensemble Monte Carlo Simulation of Complex Molecular Systems

Thomas W. Rosch and Edward J. Maginn*

Department of Chemical and Biomolecular Engineering, University of Notre Dame, 182 Fitzpatrick Hall, Notre Dame, Indiana 46556-5637, United States

Received July 19, 2010

Abstract: Acceptance rules for reaction ensemble Monte Carlo (RxMC) simulations containing classically modeled atomistic degrees of freedom are derived for complex molecular systems where insertions and deletions are achieved gradually by utilizing the continuous fractional component (CFC) method. A self-consistent manner in which to utilize statistical mechanical data contained in ideal gas free energy parameters during RxMC moves is presented. The method is tested by applying it to two previously studied systems containing intramolecular degrees of freedom: the propene metathesis reaction and methyl-*tert*-butyl-ether (MTBE) synthesis. Quantitative agreement is found between the current results and those of Keil et al. (*J. Chem. Phys.* **2005**, *122*, 164705) for the propene metathesis reaction. Differences are observed between the equilibrium concentrations of the present study and those of Lísal et al. (*AIChE J.* **2000**, *46*, 866−875) for the MTBE reaction. It is shown that most of this difference can be attributed to an incorrect formulation of the Monte Carlo acceptance rule. Efficiency gains using CFC MC as opposed to single stage molecule insertions are presented.

1. Introduction

Techniques that rely on molecular simulation to investigate systems undergoing chemical reactions can be divided into two categories. One category, ab initio methods, relies on first principle calculations to rigorously calculate energy−structure relationships. Electronic degrees of freedom are captured in these methods, which allows for direct treatment of bond breaking, distortion, and formation. These methods work well for computing the equilibrium distribution of products in the gas phase. Incorporation into time-dependent algorithms, e.g., Car−Parrinello molecular dynamics (CP-MD),[1] allows reactions to be modeled directly, in principle. The accuracy of ab initio methods depends on the level of theory. Highly accurate methods scale poorly with system size, making them very difficult to apply to the condensed phase.[2] The second category of methods involves treating the interactions between atoms and molecules through classical potentials parametrized either from quantum mechanical calculations or experimental data. These methods alleviate the need to perform computationally expensive first principle calculations at each configuration, and thus much larger systems can be studied for longer periods of time. One approach is to use "reactive" force fields that are parametrized to treat chemical bond formation and breaking directly.[3−5] While application of these reactive potentials has led to significant insight into short-time transient behavior, to date there are relatively few systems for which parameters have been developed. It has also been found that these potentials are extremely sensitive to the way in which they were parametrized.[2] A comprehensive discussion of reactive force fields and their application can be found elsewhere.[6]

A second classically based approach is to ignore transient events, such as bond formation and breakage, and focus only on the equilibrium conversion of each species. Such an approach was independently developed by Smith and Triska[7] and Johnson et al.[8] within a Monte Carlo framework. The so-called reaction ensemble Monte Carlo (RxMC) method allows reactants and products to be interconverted through a series of stochastic moves. While there is no need for potentials containing parameters that describe bond formation and breakage, RxMC does require as input ideal gas free energy differences between species, which can be obtained

* Corresponding author. E-mail: ed@nd.edu.

10.1021/ct100615j
Published on Web 01/10/2011

either from thermophysical tables or quantum mechanical calculations. Also required is a specified reaction set describing the stoichiometry of the system and relevant intermolecular potentials that accurately describe interactions in the condensed phase. In practice, RxMC is similar to grand canonical Monte Carlo[9] (GCMC), because random insertion and deletion of molecular species (during forward and reverse reaction moves) propagate the system toward equilibrium. Reaction moves eliminate any activation barriers associated with transition states or molecular diffusion, thus achieving equilibrium concentrations in a wide number of highly nonideal systems irrespective of reaction rates. RxMC has been quite successful in predicting the equilibrium behavior of reactions for many systems. A comprehensive review of the method can be found elsewhere.[2]

Most applications of RxMC have focused on small molecules where internal degrees of freedom were constrained to equilibrium values.[7,8,10−14] For these systems, there is a clean separation between the classical and quantum mechanical contributions to the Monte Carlo acceptance rule (as shown in detail below). The situation is more complicated for systems where internal degrees of freedom cannot be constrained to their equilibrium values. Keil et al.[15,16] have formulated a set of RxMC acceptance rules for linear united atom alkanes and alkenes within a conventional configurational bias Monte Carlo (CBMC) framework. As shown previously,[17−19] this type of CBMC algorithm is only valid for models without coupling between bond angles, i.e., molecules without branch points. Lísal and co-workers have also modeled systems with flexible internal degrees of freedom within the RxMC framework.[20,21] Their system contained methyl-*tert*-butyl-ether (MTBE), a molecule with a flexible dihedral angle and coupling between bond angles. It is demonstrated below that this study did not properly incorporate these classical degrees of freedom with ideal gas quantum mechanical information within their Monte Carlo acceptance rules, which results in a shift in computed equilibrium concentrations.

One of the goals of the present work is to formulate a set of general acceptance rules for RxMC that self-consistently treats quantum mechanical and classical degrees of freedom for molecules of arbitrary complexity. A second objective is to show how a biasing strategy can be utilized with RxMC to improve sampling efficiency. Previous work by Lísal and co-workers has applied the expanded ensemble method to mesoscopic reaction ensemble dissipative dynamics simulations.[22,23] The focus of this work is on an adaptive slow growth method named continuous fractional component (CFC) Monte Carlo. The rest of this paper is organized as follows: In the next section background on RxMC is provided, and acceptance rules in this work are derived. Following this, derivation of the CFC method is provided followed by simulation details for two test cases. Next, results for the test cases are presented and compared with previous works. Finally a brief summary, and a set of conclusions are provided.

2. Methods

The reaction ensemble Monte Carlo method will be discussed for a system undergoing one reaction. It is straightforward to derive the method for multiple reactions in any number of phases. Equilibrium of a single reaction involving s species is reached when the following constraint is satisfied

$$\sum_{i=1}^{s} \nu_i \mu_i = 0 \tag{1}$$

where ν_i and μ_i are the stoichiometric coefficient and the chemical potential of species i, respectively. The Hamiltonian of each molecule is assumed to be separable into quantum and classical parts,[24] such that for a given species i

$$\mathscr{H}_i = \mathscr{H}_{i,\mathrm{qm}} + \mathscr{H}_{i,\mathrm{cl}} \tag{2}$$

where $\mathscr{H}_{i,\mathrm{cl}}$ is the Hamiltonian associated with the f_i degrees of freedom one wishes to treat classically, and $\mathscr{H}_{i,\mathrm{qm}}$ is the Hamiltonian associated with the remaining degrees of freedom that are treated quantum mechanically. During the simulation, only classical degrees of freedom will be allowed to change. Equation 2 implies that the molecular partition function is separable into quantum and classical components, such that the single molecule partition function is

$$q_i = q_{i,\mathrm{qm}} q_{i,\mathrm{cl}} = \frac{q_{i,\mathrm{qm}}}{\hbar^{f_i}} \int \exp[-\beta \mathscr{H}_{i,\mathrm{cl}}] \, \mathrm{d}p_i \, \mathrm{d}r_i \tag{3}$$

where $\hbar$ is Planck's constant, $\beta = 1/k_B T$, and p_i and r_i are the momenta and generalized coordinates associated with all the classical degrees of freedom of species i. Note that the molar standard chemical potential μ_i^0 is related to the total molecular partition function by

$$\frac{\mu_i^0}{RT} = -\ln\left(\frac{q_i}{\beta P^0 \Lambda_i^3}\right) \tag{4}$$

where P^0 is the standard state pressure and Λ is the de Broglie wavelength. The molar standard chemical potential can be obtained from thermochemical property databases[25,26] or computed from gas-phase quantum mechanical calculations and will be used as an input to the RxMC acceptance rules.

The semiclassical canonical partition function for a system containing a total of s species, each species i having N_i molecules, is

$$Q(N_1, \ldots, N_s, V, T) = \prod_{i=1}^{s} \frac{q_{i,\mathrm{qm}}^{N_i}}{N_i! \hbar^{f_i N_i}} \int \exp[-\beta \mathscr{H}_{\mathrm{cl}}] \, \mathrm{d}p \, \mathrm{d}r \tag{5}$$

where $\mathscr{H}_{\mathrm{cl}}$ is the classical Hamiltonian of the system with momenta and coordinates p and r, respectively. Equation 5 implies that all intermolecular interactions are treated classically. If the classical Hamiltonian is separable into potential and kinetic contributions, then the integration of momenta in eq 5 is straightforward. Moreover, if the classical potential only involves pairwise interactions and there are no external fields, then it is possible to integrate over the translational components of the generalized coordinates. The result is that

$$Q(N_1, \ldots, N_s, V, T) = \prod_{i=1}^{s} \frac{q_{i,\mathrm{qm}}^{N_i} V^{N_i}}{N_i! \Lambda_i^{f_i N_i}} \int \exp[-\beta \mathscr{V}_{\mathrm{cl}}] \, \mathrm{d}r' \tag{6}$$

where r' represents all the classical degrees of freedom minus translational terms and $\mathcal{V}'_{\mathrm{cl}}$ is the classical pairwise potential energy associated with the classical degrees of freedom.

The volume associated with the remaining classical degrees of freedom of a given species i will be given the symbol Ω_i, such that

$$\Omega_i = \int \mathrm{d}r'_i \tag{7}$$

It follows that the volume associated with *all* the classical degrees of freedom is

$$\Omega_i V = \int \mathrm{d}r_i \tag{8}$$

The grand ensemble is the most appropriate for RxMC, and the grand partition function can be written as the canonical partition function of eq 6, expanded in chemical potential:

$$\Xi(\mu_1, ..., \mu_s, V, T) = \sum_{N_1=0}^{\infty} \cdots \sum_{N_s=0}^{\infty} \int \exp\left[\beta \sum_{i=1}^{s} N_i \mu_i - \sum_{i=1}^{s} \ln N_i! + \sum_{i=1}^{s} N_i \ln \frac{V q_{i,\mathrm{qm}}}{\Lambda_i^{f_i}} - \beta \mathcal{V}\right] \mathrm{d}r' \tag{9}$$

where the subscript cl designating the potential as classical has been dropped for simplicity.

From eq 9, the probability of state m in the grand ensemble is

$$\rho_m = \frac{1}{\Xi(\mu_1, ..., \mu_s, V, T)} \exp\left[\beta \sum_{i=1}^{s} N_{i,m} \mu_i - \sum_{i=1}^{s} \ln N_{i,m}! + \sum_{i=1}^{s} N_{i,m} \ln\left(\frac{\Omega_i V q_{i,\mathrm{qm}}}{\Lambda_i^{f_i}}\right) - \beta \mathcal{V}'_m\right] \tag{10}$$

Note that the number of each species as well as the potential energy depend upon state m.

Equation 10 will be the basis for the derivation of acceptance rules in the RxMC method. The detailed balance condition requires that moves between states m and n satisfy the following expression:[27]

$$\prod_{mn} \alpha_{mn} \rho_m = \prod_{nm} \alpha_{nm} \rho_n \tag{11}$$

where Π_{mn} is the one-step transition probability of going from state m to state n, α_{mn} is underlying matrix of the Markov chain (the move "attempt" probability in going from state m to state n), and ρ_m is given by eq 10.

Within a RxMC simulation, transitions of the system from state m to n fall into two categories. One is for transitions where no change in species composition occurs, and the second is for a reaction move where the composition does change. Let the parameter δ differentiate between the three possible cases. When $\delta = 0$, no reaction occurs. When $\delta = +1$, a forward reaction occurs such that reactants (species with a negative stoichiometric coefficient) are consumed and products (species with a positive stoichiometric coefficient) are produced. For a reverse reaction that consumes products and produces reactants, $\delta = -1$. For any species i initially in state m with $N_{i,m}$ molecules, there will be $N_{i,n}$ molecules in the new state n given by

$$N_{i,n} = (N_{i,m} + \nu_i \delta) \tag{12}$$

Combining eqs 10−12 results in the following general formulation of the one-step transition probability for RxMC moves:

$$\Pi_{mn} = \min\left(1, \frac{\alpha_{nm}}{\alpha_{mn}}\left[\prod_{i=1}^{s} \frac{N_{i,m}!}{(N_{i,m} + \nu_i \delta)!}\left(\frac{\Omega_i V q_{i,\mathrm{qm}}}{\Lambda^{f_i}}\right)^{\nu_i \delta}\right] \times \exp[-\beta(\Delta \mathcal{V}_{mn}^{\mathrm{inter}} + \Delta \mathcal{V}_{mn}^{\mathrm{intra}})]\right) \tag{13}$$

where eq 1 has been employed, and the change in potential energy has been divided into classical intramolecular and intermolecular contributions $\Delta \mathcal{V}_{mn}^{\mathrm{intra}}$ and $\Delta \mathcal{V}_{mn}^{\mathrm{inter}}$, respectively. Note that the quantum molecular partition function $q_{i,\mathrm{qm}}$ appears in eq 13 and not the total molecular partition function q_i.

If molecules are treated as rigid such that intramolecular degrees of freedom are frozen at their equilibrium values, then only intermolecular and translational degrees of freedom are treated classically (i.e., $f_i = 3$). A molecule only adopts one conformation so $q_{i,\mathrm{qm}} = q_i$, $\Delta \mathcal{V}_{mn}^{\mathrm{intra}} = 0$, and $\Omega_i = 1$. Also, if no biasing is used, then the stochastic matrix is symmetric such that $\alpha_{mn} = \alpha_{nm}$. In this case, eq 13 reduces to the conventional acceptance rule used in many previous studies in which small, rigid molecules were simulated.[7,8,10−14]

When simulations are performed on more complex molecules that cannot be treated realistically as rigid, intramolecular conformations must be modeled subject to a classical intramolecular potential function, $\Delta \mathcal{V}_{mn}^{\mathrm{intra}}$. For example, $\Delta \mathcal{V}_{mn}^{\mathrm{intra}}$ may include terms that capture bond angle bending and dihedral angle rotation. In these cases, eq 13 must be used. One must be careful to properly account for the fact that $q_{i,\mathrm{qm}}$ and not q_i appears in the acceptance rules. For example, the use of tabulated values of μ_i^0 will lead to an incorrect accounting of the classical and quantum mechanical terms (see eq 4). Below it is shown how to formulate a set of self-consistent RxMC moves that satisfy eq 13 and still allow the use of q_i or μ_i^0 in the acceptance rule.

The approach relies upon the use of a "reservoir sampling" method to generate conformations of flexible molecules with a known probability. More details of the method as well as how it can be combined with configurational biasing is found elsewhere,[17] but for simplicity only the basic method is outlined below. To apply the method, molecules are broken into fragments. Each fragment contains atoms connected only by bond lengths and angles. These degrees of freedom can be classified as "hard" because they are very strong functions of position. Each fragment is connected to another via one dihedral potential, which is a weaker function of position compared to bond lengths or angles and is thus known as a "soft" degree of freedom. This method decouples "hard" and "soft" degrees of freedom and allows for a systematic approach to build molecules that satisfy a Boltzmann distribution of internal energy. A reservoir of each kind of fragment is created via a standard Metropolis Monte Carlo

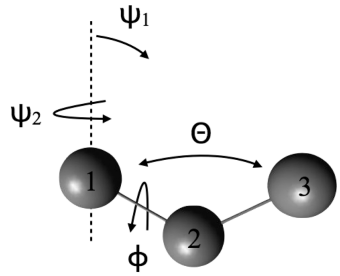


Figure 1. Internal coordinates associated with a particular three-atom fragment.

presimulation. Each fragment appears in the reservoir with a well-defined probability according to the Boltzmann weight of all the flexible degrees of freedom in that fragment. That is, the probability of a given fragment appearing in the reservoir is

$$\rho_{\text{frag}} = \frac{\exp[-\beta \mathcal{V}^{\text{frag}}]\Delta V_{\text{frag}}}{\sum_{n\text{frags}} \exp[-\beta \mathcal{V}^{\text{frag}}]\Delta V_{\text{frag}}} \approx \frac{\exp[-\beta \mathcal{V}^{\text{frag}}]\mathrm{d}V_{\text{frag}}}{\int \exp[-\beta \mathcal{V}^{\text{frag}}]\mathrm{d}V_{\text{frag}}} \tag{14}$$

where *n*frags is the number of fragments in the reservoir, $\mathcal{V}^{\text{frag}}$ is the classical potential energy of the fragment and ΔV_{frag} and $\mathrm{d}V_{\text{frag}}$ are the discrete and differential volume elements associated with the flexible classical degrees of freedom of the fragment, minus the center of mass. For example, a three-atom fragment with fixed bond lengths but a flexible bond angle, such as the one shown in Figure 1, has associated with it the translational volume of atom 1 ($\mathrm{d}V_1 = \mathrm{d}x_1\mathrm{d}y_1\mathrm{d}z_1$) and the internal coordinate volume terms. Atom 2 is specified by two Euler angles Ψ_1 and Ψ_2 as well as a Jacobian associated with Ψ_1

$$\mathrm{d}V_2 = \mathrm{d}\cos(\Psi_1)\mathrm{d}\Psi_2 = 4\pi \tag{15}$$

and for atom 3 the differential volume element is

$$\mathrm{d}V_3 = \mathrm{d}\cos(\theta)\mathrm{d}\phi = 4\pi \tag{16}$$

So in this particular case, $\mathrm{d}V_{\text{frag}} = \mathrm{d}V_2\mathrm{d}V_3$.

Each fragment required to assemble a molecule is chosen according to eq 14 and connected to one another through one bond. As will be shown below, the dihedral angle ϕ that defines the relative orientation of two fragments is chosen according to

$$\rho_\phi = \frac{\exp[-\beta \mathcal{V}'_\phi]\Delta V_\phi}{\sum \exp[-\beta \mathcal{V}'_\phi]\Delta V_\phi} \approx \frac{\exp[-\beta \mathcal{V}'_\phi]\mathrm{d}V_\phi}{\int \exp[-\beta \mathcal{V}'_\phi]\mathrm{d}V_\phi} \tag{17}$$

where $\mathcal{V}'_\phi$ contains all the energy associated with dihedral angle ϕ and any nonbonded intramolecular energy interactions between the fragments involved in the dihedral angle. For simple molecules that do not contain nonbonded intramolecular interactions, a simple rejection method prevalent in configurational bias techniques may be used to generate the correct dihedral distribution. For more complex topologies that contain nonbonded terms, a number of methods may be utilized, e.g., presimulations of single molecules in the ideal gas phase which tabulate dihedral probability functions.

Once the entire molecule is assembled, the probability of a particular conformation for a molecule *i* being inserted into the system is

$$\rho_{\text{ins},i} = \frac{\mathrm{d}V_{com}}{V} \prod_{j=1}^{n\text{frags}} \rho_{\text{frag},j}\rho_{\phi,j} \tag{18}$$

where $\mathrm{d}V_{\text{com}}/V$ accounts for the random insertion of the center of mass of the molecule. Note that

$$\mathrm{d}r' = \prod_{j=1}^{n\text{frags}} \mathrm{d}V_{\text{frag},j}\,\mathrm{d}V_{\phi,j} \tag{19}$$

When a molecule is deleted from the system, no energy bias is used. Thus the probability of a configuration of deleted species *i* is

$$\rho_{\text{del},i} = \frac{\mathrm{d}V_{\text{com}}}{V}\frac{\mathrm{d}r'_i}{\Omega_i} \tag{20}$$

The underlying Markov matrix of insertion and deletion moves can be constructed from eqs 18 and 20. The ratio of attempt probabilities is

$$\frac{\alpha_{nm}}{\alpha_{mn}} = \prod_{i=1}^{s}\left[\left(\frac{\int \exp[-\beta \mathcal{V}_i^{\text{intra}}]\mathrm{d}r'_i}{\exp[-\beta \mathcal{V}_i^{\text{intra}}]\mathrm{d}r'_i}\right)\left(\frac{\mathrm{d}r'_i}{\Omega_i}\right)\right]^{\nu_i\delta} \tag{21}$$

Since the single molecule classical partition function is

$$q_{i,\text{cl}} = \frac{V}{\Lambda^{f_i}}\int \exp[-\beta \mathcal{V}_i^{\text{intra}}]\mathrm{d}r'_i \tag{22}$$

Equation 21 becomes

$$\frac{\alpha_{nm}}{\alpha_{mn}} = \prod_{i=1}^{s}\left[\left(\frac{\Lambda^{f_i}q_{i,\text{cl}}}{\mathrm{d}r'_i V \exp[-\beta \mathcal{V}_i^{\text{intra}}]}\right)\frac{\mathrm{d}r'_i}{\Omega_i}\right]^{\nu_i\delta} \tag{23}$$

Finally, substituting eq 23 into 13, the desired result for the one-step transition probability is obtained

$$\Pi_{mn} = \min\left(1, \prod_{i=1}^{s}\left[\frac{N_{i,m}!}{(N_{i,m}+\nu\delta)!}q_i^{\nu_i\delta}\right]\exp[-\beta(\Delta \mathcal{V}_{mn}^{\text{inter}})]\right) \tag{24}$$

Note that the full single molecule partition function appears in the acceptance rule and that only the difference in intermolecular classical potential energy is used. By definition the conformations used in reaction moves already satisfy a Boltzmann distribution in regards to their intramolecular energy, and thus only intermolecular terms appear in eq 24. This acceptance rule is convenient to use because it allows one to use standard thermochemical data for the ideal gas partition function of the molecule. Note that in the absence of any intermolecular interactions, the ideal gas ratio of free energies completely determines the equilibrium concentrations of the reacting mixture, as is required. Intramolecular conformations will also appear according to an ideal gas

probability distribution since eq 23 was used. Equation 24 is the main result, but it must be emphasized that it is valid only if one generates configurations consistent with eq 23.

As mentioned in the Introduction, Lísal and co-workers modeled the MTBE synthesis reaction. MTBE contains three flexible dihedral angles that are modeled using a classical potential. From private communications with the authors, it was ascertained that they generated random configurations for MTBE such that $\alpha_{mn} = \alpha_{nm}$. To ensure that intramolecular degrees of freedom were properly sampled, the intramolecular potential energy was included in the acceptance rule. They also used the molar standard chemical potential which is related to the full molecular partition function q_i through eq 4. Thus during a reaction move, their acceptance rule was

$$\Pi_{mn,\mathrm{Lisal}} = \min\left(1, \prod_{i=1}^{s}\left[\frac{N_i^m!}{(N_i^m + \nu_i\delta)!}q_i^{\nu_i\delta}\right] \times \exp[-\beta(\Delta\mathcal{V}_{mn}^{\mathrm{inter}} + \Delta\mathcal{V}_{mn}^{\mathrm{intra}})]\right) \quad (25)$$

It is clear from eq 25 that the intramolecular contribution is counted twice; once in the full molecular partition function q_i and once in the exponential term ($\Delta\mathcal{V}_{mn}^{\mathrm{intra}}$). If one considers the simple case of an ideal gas reaction in which components contain flexible intramolecular degrees of freedom, then the acceptance probability consistent with quantum mechanics is $\Pi_{mn,\,\mathrm{Lisal}} = \min(1, \prod_{i=1}^{s}[N_i^m!/((N_i^m + \nu_i\delta)!)q_i^{\nu_i\delta}])$. Equation 25 does not reduce to this, while eq 24 does. We will show explicit simulation results supporting this claim below.

The only other atomistically detailed RxMC study the authors are aware of involving flexible intramolecular degrees of freedom was carried out by Keil and co-workers. They derived acceptance rules for RxMC within a CBMC framework to study propene metathesis within confined environments.[15,16] While their formulation is only valid for linear molecules, it is consistent with the acceptance rules given in the present work. In particular, it is easy to show that eq A11 of their work[16] reduces to eq 23 if only one trial position of inserted molecules is attempted. Results from simulations testing eqs 23 and 25 for MTBE synthesis and propene metathesis are given in the next section.

Finally, the reservoir sampling method[17] is directly suited to include configurational bias in the RxMC framework, similar to that proposed by Keil and co-workers. An alternative method to CBMC, described below, relies on slow growth to overcome large free energy barriers associated with insertion and deletion in dense fluids. In the continuous fractional component (CFC) MC method,[28] insertions and deletions are not accomplished in one step but rather by gradual changes in the intermolecular coupling of fractional molecules to the rest of the system. The coupling of the fractional molecules is controlled by a parameter λ that fluctuates between 0 and 1. At $\lambda = 0$ the reactant molecules have no intermolecular interaction with the system, while the product molecules are completely coupled to the system. At $\lambda = 1$ the opposite is true; product molecules are decoupled, while reactant molecules fully interact with the system. Regardless of the value of λ, the fractional molecules contain full intramolecular interactions (bond, angle, dihedral, improper, etc.). There is need only to define one λ associated with the reaction. Any species with a negative ν_i (reactant) will have a coupling parameter of λ, while any species with a positive ν_i (product) will have a coupling parameter of $1 - \lambda$. A major strength of slow-growth methods is their ability to incorporate a biasing function to aid in the transition through λ states. A biasing function $\eta(\lambda_j)$ that depends only on the amount of coupling between the system and the fractional molecules is used here. The semiclassical partition function for systems containing partially coupled molecules is found elsewhere.[28]

Reaction moves within a CFC framework are now replaced by moves from state m to n that consist of attempts to randomly alter the value of λ. A transition to state n where the value of the coupling parameter λ changes by an amount ξ will fall into either of two categories. The first category occurs if $0 \le (\lambda + \xi) \le 1$. In this category no addition and deletion of molecules occur, and the transition probability is

$$\prod_{mn,\lambda}^{1} = \min[1, \exp(-\beta\Delta\mathcal{V}_{mn}^{\mathrm{inter}})\exp(\eta(\lambda_n) - \eta(\lambda_m))] \quad (26)$$

Because there is no change in intramolecular energy for this category, the difference in energy in eq 26 is purely intermolecular. The biasing factors $\eta(\lambda_j)$ help overcome energy barriers and more efficiently sample λ space. Optimization of the weighting factors was done using the Wang−Landau method.[29]

The second category of λ transitions occurs if either $(\lambda + \xi) < 0$ or $(\lambda + \xi) > 1$. The former case $(\lambda + \xi < 0)$ refers to a "forward" reaction, while the latter $(\lambda + \xi > 1)$ is a "reverse" reaction. For a reverse reaction, coupling parameters of fractional reactant and product molecules are set to 1 and 0, respectively. Additionally new fractional reactant molecules are inserted into the system with a coupling parameter of $\lambda_{\mathrm{new}} = (\lambda + \xi) - 1$. Finally, random product molecules are selected from the system and their coupling parameter is set to $1 - \lambda_{\mathrm{new}}$. For a forward reaction, the coupling parameters of fractional product and reactant molecules are set to 1 and 0, respectively. New fractional reactant molecules are chosen from the system, and the coupling parameter is set to $\lambda_{\mathrm{new}} = (\lambda + \xi) + 1$. Finally, product molecules are inserted into the system with a coupling parameter of $1 - \lambda_{\mathrm{new}}$. The transition probability is very similar to that in eq 24, except that biasing values are included

$$\Pi_{mn,\lambda}^{2} = \min\left[1, \prod_{i=1}^{s}\left(\frac{N_i^m!}{(N_i^m + \nu_i\delta)!}q_i^{\nu_i\delta}\right) \times \exp(-\beta\Delta\mathcal{V}_{mn}^{\mathrm{inter}})\exp(\eta(\lambda_n) - \eta(\lambda_m))\right] \quad (27)$$

3. Simulation Details

3.1. Continuous Fractional Component Method. For simulations using the CFC MC method, a scaled potential[28]

was used to model intermolecular interactions involving fractional molecules. For Lennard-Jones (LJ) interactions the following potential was used

$$\mathcal{V}_f = \lambda_{ij} 4\varepsilon_{ij} \left\{ \frac{1}{\left[\tau(1-\lambda_{ij})^2 + \left(\frac{r_{ij}}{\sigma_{ij}}\right)^6\right]^2} - \frac{1}{\left[\tau(1-\lambda_{ij})^2 + \left(\frac{r_{ij}}{\sigma_{ij}}\right)^6\right]} \right\} \quad (28)$$

λ_{ij} was taken to be the product of the scaling parameter of each molecule, $\lambda_{ij} = \lambda_i \times \lambda_j$, and τ is an adjustable parameter that was set to 0.5 following the work of Shi and Maginn.[28] Eq 28 shows that the full LJ potential is recovered at $\lambda_{ij} = 1$. For electrostatics, the partial charges on fractional molecules were scaled as $Q_f = \lambda_i^5 Q_i$ because nonlinear scaling is known to moderate strong electrostatic interactions resulting from insertion.[28]

As described in the Methods Section, bias factors were used to help push through free energy barriers associated with insertion or deletion in a dense system. Ideally the bias factors would allow any value of the scaling parameter λ to be sampled with equal probability. Recall that the scaling parameters of reactants (λ) and products ($1 - \lambda$) are not independent, thus a bias function $\eta(\lambda_j)$ dependent only upon the coupling of reactants is defined. The Wang−Landau method[29] was found to be efficient for determining $\eta(\lambda_j)$.[28] In practice λ was divided into 10 equal intervals, [0, 0.1), [0.1, 0.2), ..., [0.9, 1], with each interval j assigned a bias factor $\eta(\lambda_j)$. Initially all bias factors were set to 0. During equilibration, after an attempt to change λ, the value of $\eta(\lambda_j)$ was modified according to

$$\eta(\lambda_j) = \eta(\lambda_j) - \upsilon \quad (29)$$

where υ is a scaling parameter, initially set to 0.01. A histogram was kept that tracked the number of times each λ interval was visited. After 10 000 attempts to change the value of λ, the histogram was checked to see if each interval was visited at least 30% as often as the most visited interval. If this criterion was satisfied, then the scaling parameter was modified according to $\upsilon = 0.5\upsilon$, and histograms were reset to 0. Once the value of υ was equal to 5×10^{-6}, $\eta(\lambda_j)$ was no longer altered.

Subsequent λ moves perturb the same molecule until that molecule becomes either fully coupled or decoupled from the simulation box. Local relaxation around this molecule is thus critical for proper and efficient sampling. Preferential sampling, introduced by Owicki,[30] was utilized so that thermal equilibration moves were attempted more often for molecules surrounding the fractional ones. Two parameters are needed for preferential sampling, the volume around a fractional molecule V_{in} in which to bias thermal equilibration moves, and a parameter that governs the percentage of thermal equilibration moves to perform within that given volume $\hat{p}$. In the present work, $V_{in} = 4/3\pi(8\ \text{Å})^3$ and $\hat{p} =$ 85%, yielding ∼50% thermal moves attempted within the preferential volume.

3.2. MTBE Synthesis. To test the present method, two different systems were studied. One was the production of MTBE from isobutene and methanol, previously studied by Lísal and co-workers.[20,21] The reaction is given by

$$C_4H_8 + CH_3OH \leftrightarrow C_5H_{12}O \quad (30)$$

This system was chosen because it is one of only two cases where RxMC has been used in the condensed phase for flexible molecules. One objective was to test how the disparity between eqs 23 and 25 affect equilibrium concentrations of reactant and products. A second objective was to compare the efficiency of the CFC MC method relative to unbiased approaches. Isobutene and methanol were modeled using a united-atom version of the "optimized potentials for liquid simulations" (OPLS) model and contained no intramolecular degrees of freedom. MTBE was also modeled using OPLS, yet it has three flexible dihedral angles. Published potential parameters[21] were used with two exceptions. First, private correspondence with the authors revealed that the σ parameter for oxygen in MTBE used in their study was 3.0 Å, not 3.8 Å as reported. Second, Lísal and co-workers treated Coulombic long-range interactions by the reaction-field method, while the Ewald summation approach[31] was used here.

All MC simulations were conducted in the isothermal−isobaric (NPT) ensemble at 360 K and 5 bar. Each simulation was initialized with 512 molecules of various methanol to isobutene ratios. Geometric mixing rules were used to calculate ε_{ij} and σ_{ij} in eq 28. The cutoff distance r_{cut} for both the LJ and real space Coulombic interactions was set to 16 Å for the condensed phase and $65 \le r_{cut} \le 75$ Å for the vapor phase. LJ long-range corrections were added to the configurational energy assuming that the radial distribution function equaled unity beyond the cutoff.[27] Ewald parameters αL and K_{max} were set to 7.0 and 7.0, respectively, for both phases. Equilibrium simulations were run for 20×10^6 MC steps, and the standard deviation of four independent production runs of 20×10^6 MC steps was taken to be the statistical uncertainty. Translations, rotations about an axis, volume changes, intramolecular rearrangements, and reaction moves were attempted with probabilities of 20, 59.9, 0.1, 10, and 10%, respectively, for simulations absent gradual insertions and deletions. For simulations utilizing the CFC MC method, translations, rotations, volume changes, intramolecular rearrangements, and λ changes were attempted with probabilities of 35, 53.9, 0.1, 10, and 1%, respectively. All simulations were performed on dual core Intel opteron processors. The 80×10^6 MC steps required 1.5 days to complete.

3.3. Propene Metathesis. The RxMC method developed in this work was also applied to a system containing multiple reactions. Propene metathesis, studied by Keil and co-workers,[15] contains the following reactions

$$2C_3H_6 \leftrightarrow C_2H_4 + \text{trans} - C_4H_8 \quad (31)$$

Table 1. Single-Phase Vapor−Liquid Equilibrium Data Computed From Gibbs Ensemble Simulations[a]

T	P^{sat}_{LSN}	P^{sat}_{RM}	phase	$\mathscr{V}_{LSN}$	$\mathscr{V}_{RM}$	V_{LSN}	V_{RM}
				Isobutene			
350	10.51_{88}	10.32_{32}	g	-1.11_{11}	-1.09_{4}	2266_{225}	2332_{96}
			l	-15.47_{18}	-15.47_{4}	111.5_{13}	111.5_{3}
320	5.13_{50}	5.21_{36}	g	-0.59_{7}	-0.60_{4}	4635_{514}	4602_{353}
			l	-16.89_{16}	-16.88_{4}	103.7_{9}	103.8_{3}
				Methanol			
450	24.07_{267}	21.53_{88}	g	-5.94_{57}	-6.22_{37}	660.8_{388}	774.0_{437}
			l	-21.39_{69}	-24.06_{9}	66.18_{365}	59.22_{68}
420	13.23_{208}	15.78_{68}	g	-5.02_{61}	-4.46_{33}	1184_{95}	1636_{46}
			l	-24.95_{33}	-26.90_{4}	54.78_{106}	52.89_{8}
				MTBE			
480	19.26_{100}	20.65_{70}	g	-2.79_{21}	-0.94_{23}	1465_{105}	1309_{101}
			l	-17.55_{36}	-15.64_{16}	170.4_{34}	167.5_{16}
440	10.08_{37}	10.61_{26}	g	-1.43_{21}	0.36_{4}	2932_{364}	2790_{47}
			l	-20.27_{27}	-18.34_{2}	150.5_{19}	150.0_{2}

[a] $\mathscr{V}$, V, P^{sat} is the potential energy in kJ/mol, molar volume in cm^3/mol, and saturated vapor pressure in bar, respectively. Subscripts LSN and RM correspond to the work of Lísal et al. and the present work, respectively. Temperatures are in Kelvin.

$$2C_3H_6 \leftrightarrow C_2H_4 + \text{cis} - C_4H_8 \quad (32)$$

$$\text{cis} - C_4H_8 \leftrightarrow \text{trans} - C_4H_8 \quad (33)$$

Only two of these reactions are independent, and therefore eq 32 was not sampled during the simulation. The species involved in the propene metathesis were modeled using the TraPPE-UA model for inter- and intramolecular interactions. The parameters can be found elsewhere.[32,33]

NPT MC simulations were performed at 5 bar and temperatures of 300, 450, and 600 K. Lorentz−Berthelot combining rules were used to calculate ε_{ij} and σ_{ij} in eq 28. LJ long-ranged interactions were added to the configurational energy for a cutoff of $r_{cut} = 80$ Å. All simulations were initialized with 800 propene molecules. Equilibrium and production runs were conducted for the same number of steps as the MTBE system. Simulations required around 1 day to complete on dual core Intel opteron processors. Because this reaction occurs in the gas phase, the CFC MC method was not used; only single stage insertion and deletions were needed. Move attempt probabilities matched that for integer insertions and deletions in the MTBE system.

4. Results and Discussion

To determine what effect, if any, calculating long-ranged electrostatic interactions via the Ewald summation method rather than the reaction field method, the Gibbs ensemble MC simulations were conducted to compute pure component vapor−liquid equilibrium. Table 1 displays the comparison between this work and that reported by Lísal et al. Isobutene equilibrium bulk properties of this work agree with that of Lísal et al. to a high statistical precision. Interestingly, the properties of methanol and MTBE agree less perfectly, although in general the results are similar. Given that the methanol and MTBE models have partial charges while the isobutene model does not, the differences may be due to the fact that Lísal et al. used the reaction field method for calculating long-ranged electrostatics, while the Ewald summation method is used in the current work. Overall the agreement of single-component bulk properties is sufficient to proceed with comparison of the reaction ensemble method.

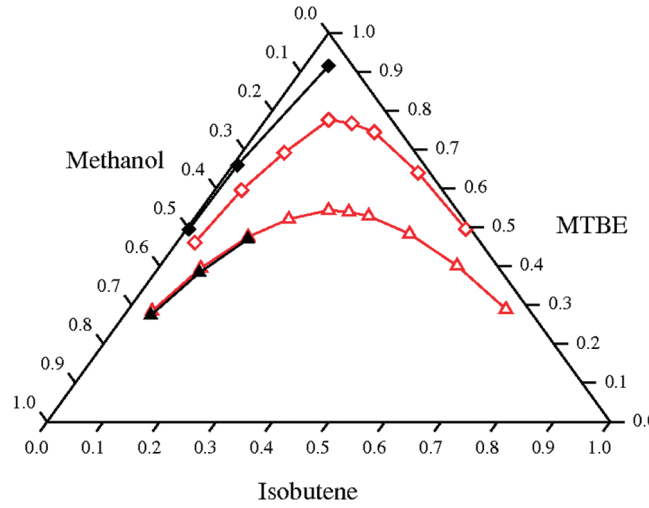


Figure 2. Triangular composition simplex for MTBE synthesis at $T = 360$ K and $P = 5$ bar. Diamonds and triangles correspond to liquid- and vapor-phase equilibrium compositions, respectively. Open symbols correspond to the work of Lísal et al.,[20] and filled symbols correspond to this work.

RxMC simulations were performed for three different initial mol ratios r of isobutene to methanol in both the gas and liquid phase at 360 K and 5 bar. Equilibrium properties were calculated in the vapor phase for $r = 0.34$, 0.51, and 0.67 and in the liquid phase for $r = 0.51$, 0.67, and 1.0. Data were first collected using single-step insertions and deletions using eq 24. A comparison of the present work with that of Lísal et al. is shown in Figure 2. Immediately apparent in the figure is how the environment affects reaction equilibria. The presence of intermolecular interactions in the condensed phase that are not found in the vapor phase shifts the reaction toward the right, resulting in more MTBE and less isobutene and methanol at equilibrium. Equilibrium compositions for the vapor phase obtained in the present study are in close agreement with the work of Lísal and co-workers, but there is a noticeable discrepancy in data for the liquid phase.

To probe the discrepancy between the liquid equilibrium composition predicted using eq 24 and that of Lísal et al., additional simulations were carried out at the same conditions in Figure 2 but using the acceptance rule of eq 25. Table 2 contains these results. Using eq 25 as the acceptance rule for a reaction move, condensed phase equilibrium mole fractions closely match the values reported by Lísal et al. However, both the liquid molar volume and vapor phase mole fractions match less well with the values of Lísal et al. Discrepancies may be due to differences in the treatment of electrostatic long-range corrections; the present work uses the Ewald summation technique, while Lísal et al. utilized the reaction field method. As this is the only difference in the method to calculate the data of the present work and that of Lísal et al., it would appear to be fortuitous that the current vapor-phase reaction results calculated using eq 24 matches the data of Lísal and co-workers. Regardless, the differences in equilibrium concentrations calculated using

Table 2. Vapor- and Liquid-Phase Equilibrium Properties for the MTBE Synthesis Reaction at $T = 360$ K and $P = 5$ bar[a]

		isobutene + methanol ↔ MTBE					
	$r =$	0.337	0.503		0.671		1.0
acc. rule	phase	gas	gas	liquid	gas	liquid	liquid
eq 24	z^1_{RM}	0.0473_{17}	0.0787_{9}	0.0045_{8}	0.1217_{37}	0.0082_{7}	0.0423_{13}
eq 25	z^1_{LSN}	0.044_{21}	0.0752_{28}	0.0319_{87}	0.1193_{47}	0.0475_{104}	0.1119_{145}
eq 25	z^1_{RM}	0.1256_{28}	0.1943_{22}	0.0232_{50}	0.2569_{39}	0.0386_{40}	0.1042_{18}
eq 24	z^2_{RM}	0.6791_{6}	0.5380_{5}	0.5008_{4}	0.4087_{25}	0.3323_{5}	0.0423_{13}
eq 25	z^2_{LSN}	0.6714_{7}	0.5295_{14}	0.5074_{44}	0.4053_{32}	0.3568_{70}	0.1119_{145}
eq 25	z^2_{RM}	0.7055_{9}	0.5959_{11}	0.5102_{25}	0.4998_{26}	0.3528_{27}	0.1042_{18}
eq 24	z^3_{RM}	0.2736_{23}	0.3833_{14}	0.4947_{12}	0.4696_{63}	0.6595_{12}	0.9153_{27}
eq 25	z^3_{LSN}	0.2846_{28}	0.3953_{42}	0.4607_{132}	0.4754_{78}	0.5957_{174}	0.7762_{290}
eq 25	z^3_{RM}	0.1689_{37}	0.2099_{33}	0.4666_{74}	0.2433_{64}	0.6086_{67}	0.7915_{36}
eq 24	$\mathscr{V}_{RM}$	-2.07_{37}	-1.64_{11}	-28.58_{12}	-0.61_{6}	-28.01_{3}	-27.32_{21}
eq 25	$\mathscr{V}_{LSN}$	-2.22_{34}	-1.57_{24}	-26.72_{51}	-1.31_{16}	-23.36_{39}	-23.67_{35}
eq 25	$\mathscr{V}_{RM}$	-2.62_{18}	-1.59_{9}	-28.10_{17}	-1.11_{9}	-27.36_{28}	-25.96_{3}
eq 24	V_{RM}	5802_{107}	5508_{35}	82.87_{28}	5684_{308}	95.00_{8}	115.2_{2}
eq 25	V_{LSN}	5152_{243}	5209_{258}	87.06_{144}	5413_{290}	99.04_{14}	117.4_{15}
eq 25	V_{RM}	4835_{170}	5388_{103}	82.12_{26}	5313_{247}	93.19_{43}	110.3_{2}

[a] Acc. rule, z^i, $\mathscr{V}$, and V correspond to the acceptance rule used, equilibrium mol fraction, molar volume, potential energy, and molar volume, respectively. Subscripts and units are the same as in Table 1.

eqs 24 and 25 clearly demonstrate the importance of a correct formulation of the acceptance rule.

To validate the equilibrium molar concentrations calculated in the present work using eq 24, satisfaction of eq 1 was tested. The total chemical potential is given by

$$\mu_i^{tot} = -k_B T \ln\left(\frac{q_i V}{\Lambda_i^3 N}\right) + \mu_i^{ex} \tag{34}$$

where μ_i^{ex} is the excess chemical potential of species i. The MTBE synthesis reaction is of the form A + B ↔ C, therefore a residual value R was defined

$$\mu_C^{tot} - \mu_B^{tot} - \mu_A^{tot} = R \tag{35}$$

Equation 34 inserted into eq 35 yields

$$-k_B T \ln\left(\frac{\dfrac{\beta P^0 V}{(N_C+1)}}{\dfrac{\beta P^0 V}{(N_B+1)}\dfrac{\beta P^0 V}{(N_A+1)}} K^0\right) + \mu_C^{ex} - \mu_B^{ex} - \mu_A^{ex} = R \tag{36}$$

where $P^0 = 1.0$ atm is the standard state pressure and $K^0 = 1.9823$ is the ideal gas equilibrium constant at 360 K. The ideal gas equilibrium constant is related to molecular partition functions through eq 4 and the following equation:

$$K^0 = \exp\left(-\frac{\sum_{i=1}^{s} \nu_i \mu_i^0}{RT}\right) \tag{37}$$

The residual in eq 36 was computed at equilibrium compositions resulting both from using eq 24 as well as eq 25 for two initial conditions: $r = 0.34$ in the vapor phase and $r = 1.0$ in the condensed phase. Values for these calculations are reported in Table 3. In the gas phase, the excess chemical potential of each species was computed using the Widom insertion method.[31] In condensed phases the Widom insertion method is known to be inefficient and at times incorrect.[34] Therefore an expanded ensemble method was used in this case.[35,36] During an RxMC simulation, 1000 equilibrium snapshots of the vapor phase were collected. Postsimulation, four independent sets of Widom insertions were performed. An independent set consisted of 1×10^5 insertions for each snapshot of the system. To calculate excess chemical potentials in the liquid-phase, four independent expanded ensemble simulations were performed for each species at the compositions of interest. Each simulation contained 15 subensembles and was run for 80×10^6 MC steps. The residuals corresponding to use of eq 24 in both vapor and liquid phases are 0 to within error, indicating that the system is in chemical equilibrium. The residuals corresponding to compositions using eq 25 statistically deviate from 0, which strengthens the argument that eq 25 is not formulated in a correct and self-consistent manner.

Table 3. Equilibrium Chemical Potentials for the Compositions of RM and LSN[a]

	isobutene(A)	methanol(B)	MTBE(C)	prefactor	R
		$r = 0.337$			
eq 24	-0.12_{1}	-0.31_{2}	-0.26_{1}	-0.28_{26}	-0.11_{26}
eq 25	-0.12_{1}	-0.39_{8}	-0.26_{1}	-1.88_{5}	-1.64_{9}
		$r = 1.0$			
eq 24	-2.11_{16}	-2.66_{4}	-4.88_{18}	-0.10_{1}	-0.21_{25}
eq 25	-3.07_{13}	-3.56_{19}	-6.33_{18}	-2.26_{1}	-1.97_{30}

[a] Each column has units of $k_B T$. The column labeled "prefactor" corresponds to the first term in eq 36.

To clear up any lingering doubts in the current method, a system devoid of any electrostatics interactions that contains intramolecular degrees of freedom was examined. Keil and co-workers[15,16] modeled propene metathesis using TraPPE-UA force fields where atomic degrees of freedom are encapsulated classically through angle and dihedral energy terms. The authors formulated acceptance rules in a similar manner to the present work to be used with configurational bias techniques and were able to reproduce experimental results with high accuracy. Conversion of propene to *cis*-butene, *trans*-butene, and ethene in bulk gas at 300, 450, and 600 K and 5 bar of pressure was calculated. The results

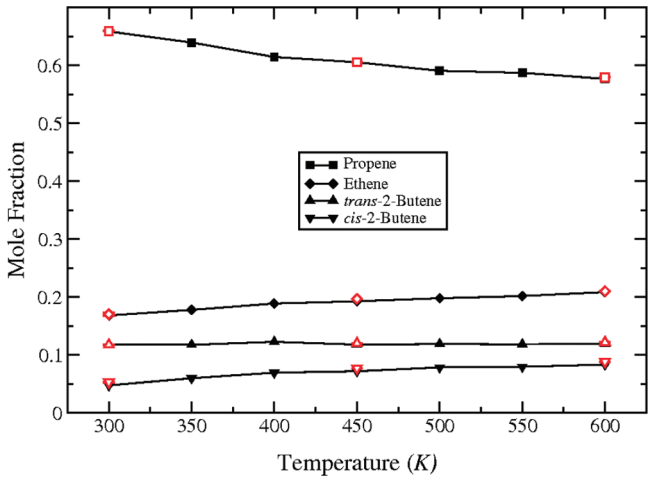


Figure 3. Equilibrium compositions for the propene metathesis reaction. Closed symbols correspond to the work of Keil et al.,[15] and open symbols correspond to the present work.

are shown in Figure 3. Results from the present work equal that of Keil and co-workers to within statistical precision. That is to say, the acceptance rule in eq 24 is consistent with the acceptance rule independently derived by Keil and co-workers. In addition, the system was simulated at 300 K using eq 25, resulting in an equilibrium mole fraction for propene of 0.89. By including the energetic penalty of creating a dihedral angle, the number of products decreased. The propene metathesis reaction at the temperatures and pressures examined in this work occurs at very dilute concentrations where one would expect mostly ideal behavior. At 300 K in an ideal gas solution, the equilibrium mole fraction of propene would be 0.657, which is very close to the results using eq 24.

Finally, the continuous fractional component method within the reaction ensemble was applied to the MTBE system. Figure 4 displays the computed equilibrium compositions using single stage insertions as well as using the CFC method. Both methods yield the same concentrations in both the condensed and gas phases. In the gas phase, though, it is unnecessary to perform gradual insertions and deletions because of the low density. Single stage reaction steps in the gas phase had an acceptance rate of ≈60%. In the liquid phase, though, single stage reaction steps had an acceptance rate <0.08%. It is for these systems, where high density causes most single stage reaction moves to be rejected, that the CFC method is expected to be most beneficial. Analysis of the efficiency of the CFC method was therefore conducted on the condensed phase MTBE synthesis reaction where the initial ratio of isobutene to methanol equaled, $r = 1.0$.

As stated in the Simulation Details Section, λ space was divided into 10 equal subsections with each given a weight, $\eta(\lambda)$, calculated during the equilibrium simulation using the Wang−Landau method.[29] These weights are inversely proportional to the free energies of the subsections, which allows the simulation to push through any free energy barrier and sample λ space equally. Figure 5 displays $-\eta$ as a function of λ for $r = 1.0$. The free energy barrier that needed to be overcome when gradually inserting a molecule was

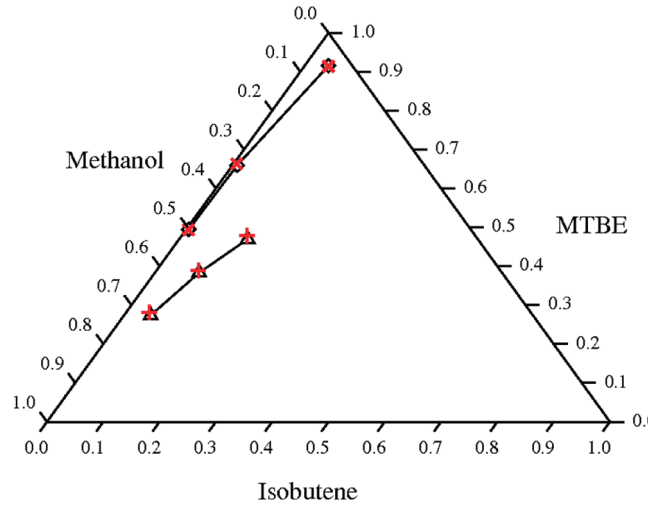


Figure 4. Triangular composition simplex for MTBE synthesis at $T = 360$ K and $P = 5$ bar. Diamonds and triangles correspond to liquid- and vapor-phase equilibrium compositions, respectively, for single stage molecule reaction moves. Plus (+) signs and × symbols correspond to vapor- and liquid-phase equilibrium compositions, respectively, for the CFC MC method.

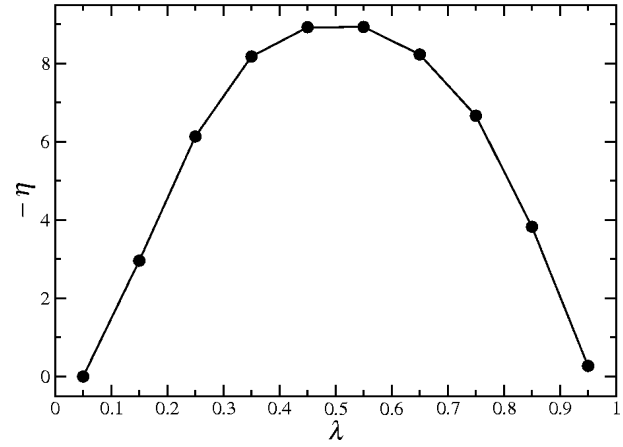


Figure 5. Inverse Wang−Landau weight for the MTBE synthesis reaction at $T = 360$ K, $P = 5$ bar, and initial isobutene to methanol mol ratio $r = 1.0$.

≈$9k_BT$, which is large enough that gradual insertion without a bias function would not be possible.

Both a CFC RxMC simulation using this weighting function and a RxMC simulation using single stage molecule insertions and deletions were run for the same amount of time. Figure 6 displays the number of isobutene molecules in the simulation box as a function of time for both methods.The averages of the CFC and the integer method were calculated to be $N_I^{avg} = 0.041 \pm 0.007$ and 0.044 ± 0.010, respectively. Overall both methods gave statistically equivalent equilibrium concentrations, yet the integer method's uncertainty was larger. This is a result of fluctuations associated with integer molecule insertions and deletions being larger than those for CFC.

While the use of the CFC method results in the same equilibrium concentration as single stage molecule insertions and deletions, the potential benefit of the method is its efficiency when insertions are difficult. The single stage

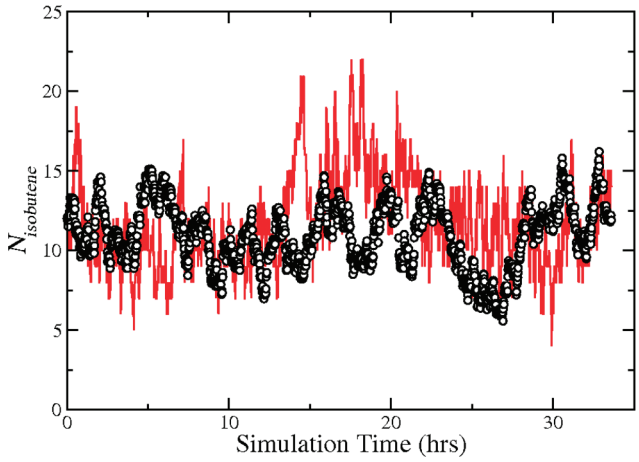


Figure 6. Number of isobutene molecules as a function of simulation time. The solid line and open symbols correspond to integer molecule insertion and deletion and CFC, respectively. For the CFC MC method $N_{\text{isobutene}} = N_{\text{isobutene}}^{\text{integer}} + \lambda$.

method achieved acceptance rates under 0.1%. Low acceptance rates mean that much of the time during simulation the system does not actually change configuration or density. The slow growth method allowed for the density to change gradually and thus more frequently. For all CFC simulations a maximum $\Delta\lambda$ of 0.3 was used. For λ moves of the first category (no additions or deletions of molecules) the acceptance rate was ≈35%. For changes in λ that resulted in new fractional molecules inserted into the system the acceptance rate was a little lower at ≈15%. These acceptance rates change the system configuration and density much more frequently than the integer insertion method, yet still may not be the best metric to show the efficiency of CFC MC. Even though reaction moves may be accepted, they may result in fluctuations about an integer molecule number, i.e., from $N_1 = 14.9 \rightarrow N_1 = 15.05 \rightarrow N_1 = 14.97$. Two accepted reaction moves of this type are not comparable to two accepted reaction moves with the integer method. Therefore whole number molecule changes within CFC MC calculations were tracked. A whole number change occurred, for example, for a system starting at $N_1 = 14.9$ only if the system changed to $N_1 < 14.0$ or $N_1 > 16.0$. Using this metric for the simulation corresponding to Figure 6, the integer method results in 4283 full molecule changes, while the CFC method results in 5083 (18.7% greater). Please note that no effort to optimize the parameters associated with the CFC method (maximum $\Delta\lambda$, number of subensembles, attempt probability for λ moves, V_{in} and $\hat{p}$ for preferential bias) has been made, which may further increase its efficiency. Also it is important to note that the increase of whole number molecule changes occurred despite changes in λ being attempted 10 times less often than in reactions using single stage molecule insertions (see Simulation Details Section).

5. Conclusions

Acceptance rules have been developed for the reaction ensemble that enables the simulation of molecules of arbitrary complexity with flexible intramolecular degrees of freedom. The acceptance rules have been developed using both a single stage transformation method as well as a "slow growth" staged deletion and insertion procedure designed to make transitions more efficient for large complex molecules. The approach was tested by simulating two systems previously examined with RxMC: MTBE synthesis and the propene metathesis reaction.

For the MTBE system, differences in composition were observed between the present work and that of Lísal et al.[20,21] It was shown that most of the discrepancies were due to a small difference in the acceptance rules used in the two studies. It is argued that the acceptance rule developed in the present work should be used when molecules with flexible intramolecular degrees of freedom are simulated.

For the propene metathesis reaction, the results obtained in the present study agree quantitatively with those of Keil and coworked.[15] Their formulation of acceptance rules within the configurational bias sampling scheme is formally identical to those developed in the present work.

The use of the slow growth continuous fractional component Monte Carlo method improved computational and sampling efficiency as compared to single stage molecule insertions and deletions. While in both cases it was possible to carry out the simulations without the use of a slow growth method, it is anticipated that for larger molecules or those in either a highly confined environment or a very dense phase, this type of slow growth approach will be essential.

Acknowledgment. The authors thank Martin Lísal and William Smith for discussions and suggestions that were critical to this work. Computational resources were provided by the University of Notre Dame Center for Research Computing. This material was financially supported by the Department of Energy (National Energy Technology Laboratory) under award number DE-FC26-07NT43091.

References

(1) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(2) Turner, C. H.; Brennan, J. K.; Lisal, M.; Smith, W. R.; Johnson, J. K.; Gubbins, K. E. *Mol. Simul.* **2008**, *34*, 119–146.

(3) Brenner, D. W. *Phys. Rev. B: Condens. Matter* **1990**, *42*, 9458–9471.

(4) Stuart, S. J.; Tutein, A. B.; Harrison, J. A. *J. Chem. Phys.* **2000**, *112*, 6472–6486.

(5) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. *J. Phys. Chem. A* **2001**, *105*, 9396–9409.

(6) Santiso, E. E.; Gubbins, K. E. *Mol. Simul.* **2004**, *30*, 699–748.

(7) Smith, W. R.; Triska, B. *J. Chem. Phys.* **1994**, *100*, 3019–3027.

(8) Johnson, J. K.; Panagiotopoulos, A. Z.; Gubbins, K. E. *Mol. Phys.* **1994**, *81*, 717–733.

(9) Norman, G. E.; Filinov, V. S. *High Temp.* **1969**, *7*, 216–222.

(10) Turner, C. H.; Johnson, J. K.; Gubbins, K. E. *J. Chem. Phys.* **2001**, *114*, 1851–1859.

(11) Carrero-Mantilla, J.; Llano-Restrepo, M. *Fluid Phase Equilib.* **2004**, *219*, 181–193.

(12) Carrero-Mantilla, J.; Llano-Restrepo, M. *Fluid Phase Equilib.* **2006**, *242*, 189–203.

(13) Lisal, M.; Nezbeda, I.; Smith, W. R. *J. Chem. Phys.* **1999**, *110*, 8597–8604.

(14) Lisal, M.; Brennan, J. K.; Smith, W. R. *J. Chem. Phys.* **2006**, *124*, 064712.

(15) Hansen, N.; Jakobtorweihen, S.; Keil, F. J. *J. Chem. Phys.* **2005**, *122*, 164705.

(16) Jakobtorweihen, S.; Hansen, N.; Keil, F. J. *J. Chem. Phys.* **2006**, *125*, 224709.

(17) Macedonia, M. D.; Maginn, E. J. *Mol. Phys.* **1999**, *96*, 1375–1390.

(18) Vlugt, T. J. H.; Krishna, R.; Smit, B. *J. Phys. Chem. B* **1999**, *103*, 1102–1118.

(19) Martin, M. G.; Siepmann, J. I. *J. Phys. Chem. B* **1999**, *103*, 4508–4517.

(20) Lisal, M.; Smith, W. R.; Nezbeda, I. *AIChE J.* **2000**, *46*, 866–875.

(21) Lisal, M.; Smith, W. R.; Nezbeda, I. *J. Phys. Chem. B* **1999**, *103*, 10496–10505.

(22) Lisal, M.; Brennan, J. K.; Smith, W. R. *J. Chem. Phys.* **2006**, *125*, 164905.

(23) Lisal, M.; Brennan, J. K.; Smith, W. R. *J. Chem. Phys.* **2009**, *130*, 104902.

(24) Hill, T. L. *Statistical Mechanics: Principles and Selected Applications*; Dover Publications: New York, 1987.

(25) Pedley, J. *Thermodynamical Data and Structures of Organic Compounds, TRC Data Series*; Thermodynamic Research Center: College Station, TX, 1994.

(26) Chase, M. *NIST-JANAF Thermochemical Tables*, Journal of Physical and Chemical Reference Data, Monograph 9; American Physical Society: Melville, NY, 1998.

(27) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.

(28) Shi, W.; Maginn, E. J. *J. Chem. Theory Comput.* **2007**, *3*, 1451–1463.

(29) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

(30) Owicki, J. C.; Scheraga, H. A. *Chem. Phys. Lett.* **1977**, *47*, 600–602.

(31) Frenkel, D.; Smit, B. *Understanding Molecular Simulation, From Algorithms to Applications*; Academic Press: New York, 2002.

(32) Martin, M. G.; Siepmann, J. I. *J. Phys. Chem. B* **1998**, *102*, 2569–2577.

(33) Wick, C. D.; Martin, M. G.; Siepmann, J. I. *J. Phys. Chem. B* **2000**, *104*, 8008–8016.

(34) Kofke, D. A. *Mol. Phys.* **2004**, *102*, 405–420.

(35) Shah, J. K.; Maginn, E. J. *J. Phys. Chem. B* **2005**, *109*, 10395–10405.

(36) Paluch, A. S.; Jayaraman, S.; Shah, J. K.; Maginn, E. J. *J. Chem. Phys.* **2010**, *133*, 124504.

CT100615J

JCTC Journal of Chemical Theory and Computation

Estimation and Inference of Diffusion Coefficients in Complex Biomolecular Environments

Christopher P. Calderon*,‡

Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005-1892, United States

Received September 9, 2010

Abstract: The 1-D diffusion coefficient associated with a charged atom fluctuating in an ion-channel binding pocket is statistically analyzed. More specifically, unconstrained and constrained molecular dynamics simulations of potassium in gramicidin A are studied. Time domain transition density based inference methods are used to fit simple stochastic differential equations and also to carry out frequentist goodness of fit tests. Particular attention is paid to varying the time between adjacent time series observations due to the well-known "non-Markovian noise" that can appear in this system due to inertia and other unresolved coordinates influencing the dynamics. Different types of non-Markovian noise are shown by the goodness of fit tests to be statistically significant on vastly different time scales. On intermediate scales, a Markovian model is not rejected by the tests; models calibrated at these intermediate scales demonstrate a predictive capability for some physical quantities. However, in this intermediate regime, ergodic sampling does not occur over the length of a time series, but a *local* diffusion coefficient is deemed statistically acceptable for the observed raw data. It is demonstrated that a linear mixed effects model can be used to summarize the variation induced by slow unresolved degrees of freedom acting as a non-Markovian noise source. The utility of quantitative criteria for assessing low-dimensional stochastic models calibrated from time series generated by high-dimensional biomolecular systems is briefly discussed. Less coarse-grained data summaries of this type show promise for better understanding the kinetic signature of unresolved degrees of freedom in time series coming from simulations and single-molecule experiments.

1. Introduction

Computations of the effective diffusion coefficient associated with a given order parameter are of interest for various reasons in complex biological systems.[1−6] For example, if the assumptions required by transition state type theories are met, then this information can be used to estimate mean first passage times. Another use of the effective (local or global) diffusion coefficient is in summarizing the statistics of unresolved forces in simulations and experiments.[7] The ability to quantify unresolved forces is particularly relevant to single-molecule experiments where interesting events occur on time scales below the temporal resolution of the measurement device.[7−11] However, in both experiments and simulations, several factors complicate unambiguous estimation of the diffusion coefficient, e.g., inertial effects,[12] measurement apparatus noise,[13,15] and nonergodic sampling of phase space.[14,16,17] Artifacts of these types of factors are sometimes reflected in a dependence of the estimated diffusion coefficient on the spacing between adjacent observations.[3,12]

In this study, simulations of a potassium ion diffusing in the binding pocket of a narrow ion channel, gramicidin A (gA), are analyzed. The ion is allowed to fluctuate in the primary binding pocket of the channel with and without external forces influencing the dynamics. Stochastic differential equation (SDE) models are fit to time series coming from these simulations using time domain methods.[18] The gA system is well-studied[1,4,17,19−22] and is of interest due

* E-mail: Chris.Calderon@numerica.us.

‡ Present Address: Numerica Corporation, Loveland, Colorado 80538, United States.

10.1021/ct1004966
Published on Web 01/07/2011

to the fact that a "memory kernel " is measurable on O(fs−ps) time scales. However, solvation effects, channel undulations, and other phenomena occurring at a broad range of time scales complicate estimating a single global diffusion coefficient from O(ns) times series.[17,18,22] Particular attention is paid to the dependence of the global diffusion coefficient on the time between adjacent observations (this time is known as the "subsampling" or "downsampling" parameter[23−25]) and on the dependence of the local diffusion coefficient on initial conditions. In cases where the latter effect is found to be statistically significant, mixed effects models[26,27] are used to provide a less coarse-grained description of the data. In all cases, goodness of fit tests[12,28] are used to assess the suitability of using a 1-D SDE to describe data arising from a high dimensional complex system. Beyond demonstrating methods that provide quantitive summaries of noisy trajectories, it is also shown how the goodness of fit tests can be utilized to help in determining which models will have predictive power for quantities of interest (such as the sum of squared displacements vs time). The techniques shown are also applicable to experimental time series where "thermal" and instrument noise exist.[13,15] The basic motivation is to efficiently infer information from experimentally accessible quantities (like force and position time series) generated by a complex system where many other degrees of freedom are not directly resolved but their influence may be detected indirectly by kinetic signatures contained in the data.[12−14,17,29]

The remainder of the article is organized as follows: Section 2 reviews the SDE model, summarizes the salient features of the statistical tools used, and provides the molecular dynamics (MD) simulation details. Section 3 presents the results and discussion, and section 4 concludes the article. Supporting Information containing additional mathematical details, a descriptive outline of the fitting procedure, and additional plots are available online.

2. Background and Methods

Computing the asymptotic slope of the mean square displacement of a freely diffusing tagged particle in a homogeneous medium plotted against time is one classic approach to defining the diffusion coefficient. An equation summarizing this idea reads

$$D \equiv \lim_{\Delta t \to \infty} \frac{\langle (z(\Delta t) - z(0))^2 \rangle}{2\Delta t} \tag{1}$$

where D denotes the "classic" global diffusion coefficient, z represents the order parameter (here, the position of the molecule evolving in 1-D), angled brackets denote an ensemble average, and Δt is the time elapsed since the initial observation $z(0)$. Ergodic sampling is usually explicitly or implicitly assumed.[24,30] There are several complications associated with applying this approach to time series coming from biomolecules; e.g., z is often confined by a nontrivial potential, there are unresolved degrees of freedom which make the dynamics "non-Markovian", the medium is not homogeneous, ergodic sampling is difficult to ensure, etc. Several approaches in the physical sciences have attempted to deal with some of these complications.[3,4,30,31] For example, under the assumption of stationary ergodic sampling of phase space, one can utilize the autocorrelation function along with an estimate of the "instantaneous variance" of the observable being monitored to obtain an estimate of the global diffusion coefficient.[3,30,31] However, rigorous unambiguous statistical methods for testing the potential sources of model misspecification given time series data and an assumed continuous time stochastic model are often not employed;[32] this issue will be expanded on in section 3.

An alternative approach to quantifying the fluctuations and computing the "local diffusion coefficient" is to use likelihood-based techniques. For simplicity, an Ornstein−Uhlenbeck SDE is considered in this article as a surrogate for the dynamics. In statistical physics, this SDE is often written as

$$\begin{aligned} \frac{dz}{dt} &= \kappa(\alpha - z) + \eta_t \\ \langle \eta_t \eta_s \rangle &= \delta(t - s) 2\tilde{D} \end{aligned} \tag{2}$$

where here the *local* diffusion coefficient is denoted by $\tilde{D}$. The tilde is used to emphasize that this is a local diffusion coefficient associated with a given SDE. η_t represents the value taken by a mean zero Gaussian process drawn at time t, and $\delta(...)$ is the Dirac δ function, which is meant to suggest that the "random force" increments are statistically uncorrelated. The parameters α and κ can be interpreted as the process mean and effective spring constant, respectively. Defining $\tilde{D}$ *does not* necessarily require one to appeal to ensemble quantities (such as a stationary autocorrelation function[3,4]) of the system observable(s), as is often the case with D. The value of $\tilde{D}$ can be estimated along individual trajectories which may not "ergodically" explore phase space.[12,14,17,33,34] However, physically interpreting $\tilde{D}$ can require care even if the model is judged statistically acceptable. $\tilde{D}$ can potentially contain signatures of unresolved degrees of freedom.[14,17] Several methods for estimating $\tilde{D}$ and quantitatively assessing various assumptions behind a proposed 1-D SDE calibrated from time series of more complex processes (e.g., in our case, the data are generated by a high-dimensional MD simulation) have recently appeared in the mathematical statistics and stochastic processes communities; some examples can be found in refs 18, 23, 35, and 36. In this body of literature, the SDE in eq 2 is often denoted by the following:

$$dz_t = \kappa(\alpha - z_t)dt + \sqrt{2\tilde{D}}dB_t \tag{3}$$

where B_t represents the standard Brownian motion process,[37] the subscript denotes the time index, and $\theta \equiv (\alpha,\kappa,\tilde{D})$ is a vector denoting the parameters needed to specify the process. [All stochastic integrals and SDEs used for modeling are of the Itô type.] For a given discretely observed time series, $\{z_i\}_{i=0}^{N}$, the maximum likelihood estimate (MLE) of the parameter, denoted by $\hat{\theta}$, can be found explicitly for the SDE in eq 3.[18] A guideline outlining some basic recommendations for fitting more general SDEs to trajectories can be found in the Supporting Information. A major advantage of utilizing modern SDE inference tools[18,23,35−37] is that unambiguous

statistical quantities can be computed and various assumptions behind a proposed surrogate SDE (possibly more involved than eq 3) evolution equation can be tested given data arising from a high-dimensional biomolecular system.[12−28] To illustrate the relevance of such statistical inference tools, consider the following: In the narrow gramicidin A channel studied in this article, it is known that inertial memory can complicate using a simple SDE like that given in eq 3 for accurately approximating/summarizing the dynamics of how an ion diffuses along the axis of the channel. If data are sampled every femtosecond, the complex statistical (temporal) dependence in the time series $\{z_i\}_{i=0}^{N}$ would not permit an SDE driven by a standard Brownian motion to approximate the dynamics. In an attempt to "average out" short time non-Markovian noise and attempt to estimate an SDE a statistically acceptable proxy, one can introduce a parameter, n, which subsamples (a.k.a. downsamples) observations.[23−25,33] For example, one can use the series $\{z_{i\times n}\}_{i=0}^{N\times n}$ to obtain $\hat{\theta}^{(n)}$, where the superscript stresses the subsampling parameter. As n increases, the influence of inertia and other fast scale motion decreases, and a process driven by Brownian motion becomes intuitively more plausible.[12]

One set of results in this article focuses on varying n and using the data coming from a high-dimensional biomolecular simulation to determine the goodness of fit of a simple SDE model. For frequently sampled data (corresponding to low n), the results are as expected; in the case of "coarsely" sampled data (corresponding to high n), slow scale unresolved motions will be shown to complicate the use of an SDE of the form given in eq 3 (or 2) to approximate statistics of the underlying complex system. Note that $N \times n$ is selected to be relatively small compared to the total time series size generated by the MD simulation. Hypothesis testing machinery, with adequate power in small samples, is useful for quantitatively determining when a simple Markovian SDE governing the dynamics is statistically acceptable given data coming from a more complex system.

The primary mathematical equations utilized in the statistical analysis are deferred to the Supporting Information; however, the basic idea behind the goodness of fit test statistics is sketched here. The time series arising from the system of interest, $\{z_{i\times n}\}_{i=0}^{N\times n}$, likely possess nontrivial temporal dependencies.[14,17,33,34] Carrying out goodness of fit tests that reliably check for temporal dependencies not implied by the assumed surrogate model class can be problematic.[28,38,39] However, if the data generating process is posited explicitly, it is possible to introduce a transformation utilizing information about all moments assumed by the proposed model which maps a correlated, stationary or nonstationary, time series to a new series of random variables, $\{Z_{i\times n}\}_{i=0}^{N\times n}$ (the transformed series is denoted by a capital letter). Under correct model specification, the Z_i's are independent and identically distributed (iid) random variables with a uniform $U[0,1]$ distribution regardless of the dependence structure.[35] The transformation with these properties does not require asymptotic arguments, and hypothesis tests can be established which simultaneously check if the transformed Z_i's are iid and have the $U[0,1]$ shape. [However, it is emphasized that the transformation requires that the data generating process be exactly known for the precise results to hold; if a parameter(s) needs to be estimated from data, then some technical complications are encountered.[35,38]]. Deviance from either condition suggests the surrogate SDE is not faithful to the data. Hong and Li proposed the so-called "omnibus" Q test statistic (relevant equations reported in Supporting Information) which jointly checks both the iid and $U[0,1]$ shape assumption. Such "omnibus" tests can sacrifice some power,[40] but tests which focus more on the independence assumption (and loosely "focus" on non-Markovian errors) can be employed. The M test in ref 35 is one such test. It does not check for the $U[0,1]$ shape but instead focuses on autocorrelations in moments of the Z_i's. The utility of both test statistics in analyzing time series possessing noise coming from many time scales will be presented.

It should be stressed that increasing n does not necessarily guarantee that a single SDE of the type given in eq 3 will provide a statistically acceptable model of the stochastic dynamics of an ion in a binding pocket. This will be demonstrated explicitly in the Results and Discussion section. For cases where ergodic sampling does not occur, but a local SDE model is deemed appropriate by some criterion at a given time scale, quantification of the influence of initial conditions on the estimated local diffusion is of interest. To accomplish this, the framework of mixed effects models will be used.[26] The setup is as follows:

$$\begin{gathered}\tilde{D}_{i,j} = \mu^{\tilde{D}} + b_i^{\tilde{D}} + \varepsilon_{i,j} \\ i = 1, 2, ..., N_{\mathrm{IC}} \\ j = 1, 2, ..., N_{\mathrm{Rep}}\end{gathered} \tag{4}$$

The more technical details of the model are deferred to the Supporting Information; the physical motivation for the terms above is described here. $\mu^{\tilde{D}}$ denotes a fixed-effect population mean of the local diffusion coefficients, and $b_i^{\tilde{D}}$ denotes a random-effect specific to initial condition i. N_{IC} represents the number of initial configurations (ICs) analyzed, where an IC is defined by the position of all atoms in a simulation. N_{Rep} denotes the number of repeat experiments for a given fixed IC (the position of all atoms is fixed, but different initial velocities are used), and $\varepsilon_{i,j}$ represents "sampling noise". The significance of the random term (i.e., dependence on initial conditions) is tested using both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Because ergodic sampling may not occur, the random-effect, i.e., variation due to the initial conditions, may be statistically significant.

2.1. Simulation Details. The NAMD[41] simulation package with parameters used originally in ref 5 and then in ref 17 is used. The temperature was set to 310 K, the pressure was maintained at 1 atm. The only significant difference in the simulations reported here is that the harmonic guiding potential is not used to "steer" the ion in a time-dependent fashion. A configuration where a single potassium ion was located in the binding pocket was used as an initial condition. This initial configuration was equilibrated for 1 ns of simulation time (without external force). After this equilibration, an ensemble of production runs carried out for 6 ns

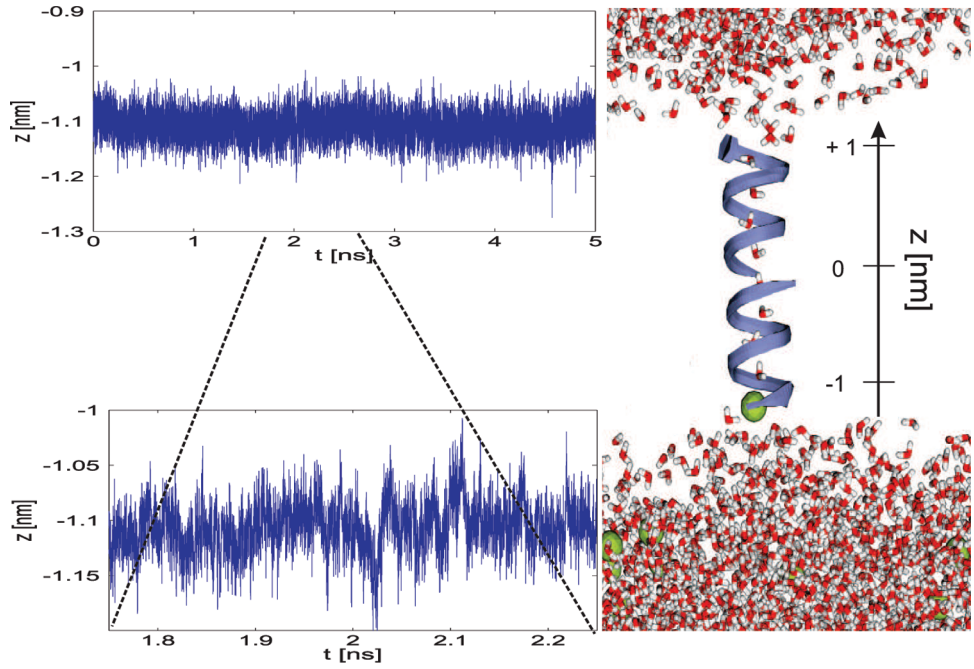


Figure 1. Sample trajectories coming from MD simulation. The bottom panel zooms in on the top panel time series to emphasize that the local mean changes in a nontrivial fashion (i.e., the potential is not a single harmonic well). In this article, we focus on the binding pocket near −1 nm. The system snapshot was generated with VMD.[48]

was used to generate additional "equilibrated" initial conditions (the tagged ion remained in the binding pocket for this time period). Every 10 fs, the position of the ion along the channel was output to disk; this sampling frequency corresponds to the parameter $n = 1$ in the expression $\{z_{i\times n}\}_{i=0}^{N\times n}$. The constrained runs used the same initial conditions as the unconstrained runs.

3. Results and Discussion

A representative trajectory obtained while monitoring the ion's position along the axis channel is plotted in the top left panel of Figure 1; the bottom left panel zooms in on a segment to show the fine temporal structure. The illustration to the right is a snapshot of the channel (lipid molecules omitted from the plot for clarity). The center of the dimer channel defines the zero of the z coordinate. The binding pocket near $z \approx -1$ nm is relatively shallow (it is computed to be $\approx 5k_BT$) but deep enough to allow the ion to be trapped for a substantial amount of MD simulation time. The 1-D potential of mean force (see Supporting Information Figure 1) used to describe z is clearly not a perfect harmonic potential nor does the PMF capture all of the information needed to understand the rich dynamics,[17,20,22] but it does describe the average location of a trapped ion fairly well over a 1−9 ns window. Note that the primary interest throughout is in the unconstrained case, but simulations utilizing a harmonic guiding potential are also studied to demonstrate that the findings are not solely an artifact of an anharmonic potential.

Figure 2 plots $\tilde{D}$ as a function of n for time series batches each consisting of $N = 100$ observations (in this plot, a total of 6×10^5 time series entries were analyzed). If n increases, while at the same time the number of (uniformly sampled) time series observations, N, is fixed, this clearly implies that

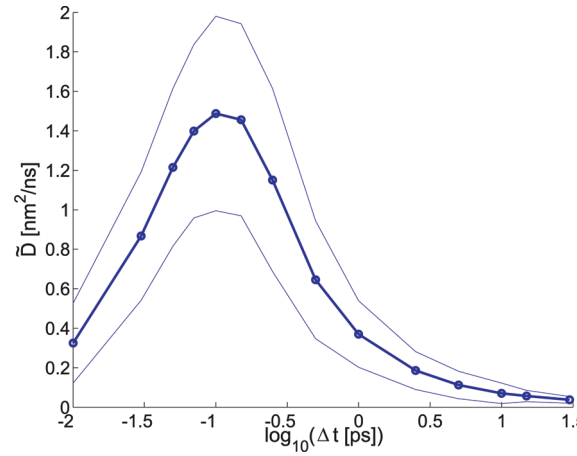


Figure 2. Estimated local diffusion coefficient for different "downsampling or subsampling" intervals. The average MLE computed from a time series appears as a symbol for a given Δt value. The solid line represents the sample mean ±2 times the sample standard. It should be noted that every observation was utilized, similar to the approach used in ref 23, but correlation in the time series does make these confidence intervals suspect. However, the trends observed in the MLE confidence bands computed are similar to those expected asymptotically if the underlying SDE was the data generating process.[18]

the time series batches constructed using a larger n are associated with a larger time between observations (and also a larger final time horizon). The logarithm of the time between adjacent entries of the (uniformly spaced) time series, denoted by $\Delta t \equiv n\delta t$, serves as the x axis in Figure 2. δt corresponds to the spacing associated with $n = 1$ ($\delta t = 10$ fs throughout). The local effective diffusion coefficient, $\tilde{D}$, estimated in this fashion displays a nontrivial trend with the "coarse-graining" parameter Δt. Dependence of $\tilde{D}$ on Δt

can either be a sign of "memory" due to inertial effects[4] and/or a sign of a "poor reaction coordinate".[2] Note that for larger Δt, $\tilde{D}$ appears to level off to a value of $\approx 3.0 \pm 1.9$ Å^2/ns ($\approx 0.030 \pm 0.019$ nm^2/ns); this "convergence" may seem to suggest that diffusive motion is an adequate approximation of the dynamics on this coarser time scale. It is interesting to also observe that the *global* diffusion coefficient estimated using a method[3] appealing to an integration of the empirically measured autocorrelation (obtained using a 9 ns MD trajectory sampled every 10 fs) and an empirical estimate of the variance is 2.7 Å^2/ns. This is in agreement with the range predicted by the MLE estimate of $\tilde{D}$ obtained using a simple Ornstein−Uhlenbeck SDE. Autocorrelation and/or memory is used traditionally in molecular dynamics.[3,4,32] Such methods usually implicitly assume that various moments are stationary and adequately sampled.

The apparent convergence of $\tilde{D}$ at larger Δt and the consistency of the local diffusion coefficient and the global diffusion coefficient (estimated using vastly different methods) might lead one to conclude that this diffusion coefficient is a reasonable summary of the dynamics which can be used for predictive purposes (such as computing the sum of squares of increments or a mean first passage time). The strong dependence of $\tilde{D}$ on the Δt for Δt in the $\approx 0.05-0.20$ ps range would also seem to suggest that this diffusion coefficient is physically meaningless (or at least nontrivial to interpret in terms of classical statistical mechanics).

However, the results shown in Figure 3 provide results suggesting that the above intuition is misleading. Here, results obtained by computing the "Q" and "M" goodness of fit tests reported in ref 35 (the relevant equations are reproduced in the Supporting Information) are plotted using the MLE obtained using observational data and the assumed SDE model hoping to approximate the high dimensional process generating the time series. If the model is correct, then it can be shown that both statistics asymptotically converge to mean zero standard normals.[35] Some simple finite sample size correction to the test statistic distribution[28] can be made (one is discussed in the caption of Figure 3); more sophisticated approaches are discussed in ref 38. Models inconsistent with the observed data (i.e., model mis-specification) will result in large values of the test statistic if there is enough statistical evidence of model inadequacy. Recall that the omnibus Q test aims to simultaneously check that the increments of the discretely observed time series follow the expected distribution shape *and* have the correlation structure consistent with the assumed model. In the surrogate SDEs considered in this paper, the dynamics of z need to be approximately Markovian for the assumed proxy to be statistically acceptable. Somewhat surprisingly, the most plausible Markovian SDE (as judged by *both* the M and Q) test statistic occurs at an intermediate Δt. Various items related to this observation are explored in the results that follow.

Figure 4 plots the empirically determined autocorrelation (AC) function of the force and position taken from three different MD simulations each spanning 3 ns. The force AC demonstrates an oscillatory behavior that decays after a fairly

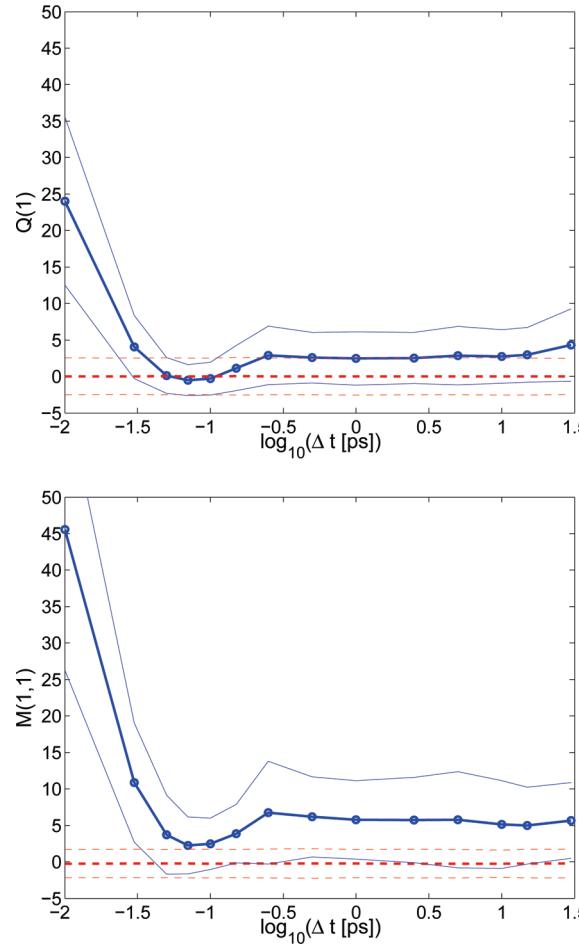


Figure 3. Goodness of fit tests. Similar to the previous figure except the corresponding M and Q statistics are plotted. The dotted lines denote the sample mean ±2 times the sample standard deviation of the test statistic one can expect under a correctly specified model. The dotted lines were generated using an iid $U[0,1]$ random variable sequence possessing the same size as the time series and evaluating Q and M with this reference sequence; the iid $U[0,1]$ reference sequence used the same nonparametric estimators as the MD data and the corresponding computed generalized residuals.

short MD timespan (≈ 0.5 ps). This nontrivial, but quickly decaying, AC in the force is the motivation for introducing a "memory kernel" in studies analyzing this channel, e.g., ref 4. It is worth noting that in each of the ACs there is a point near 0.1 ps corresponding to zero temporal correlation. This point also corresponds to the Δt (or n) where a Markovian SDE provides the best fit in regards to the goodness of fit test statistics studied. Loosely speaking, the assumptions that inertia can be ignored and the net effect of unresolved degrees of freedom can be modeled as a mean zero "random bath force" which can be approximated by a Brownian motion process become most plausible in this regime. This also might explain why the Q test has less power than the M test in this situation: the former focuses on the shape and statistical dependence between the generalized residuals computed from z_i and $z_{i\times n}$, whereas the M statistic focuses on the full AC of moments of the generalized residuals computed from the observed data and the assumed model. If inertia and other unresolved forces were truly

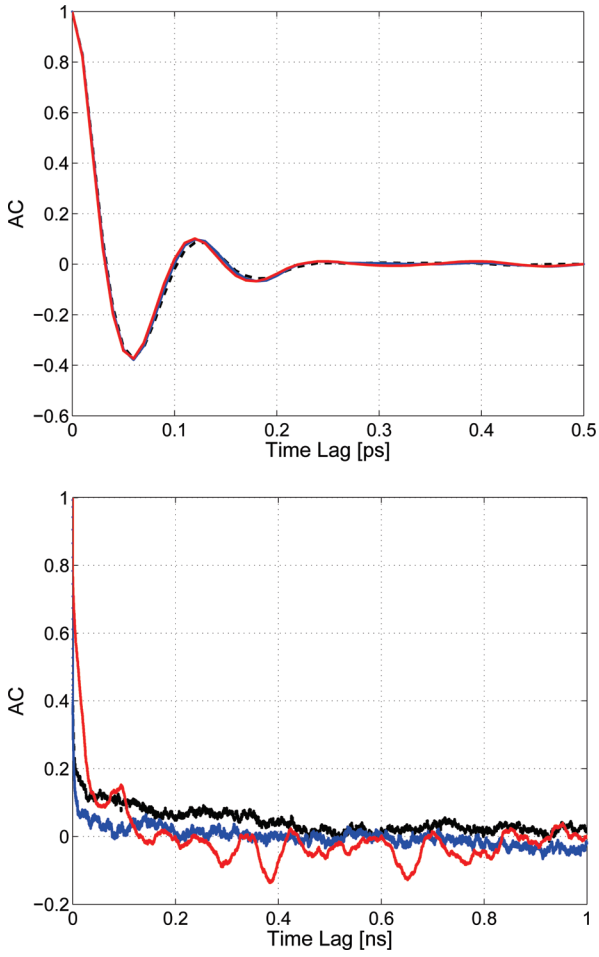


Figure 4. The measured force (top) and position (bottom) autocorrelation (AC) functions. Three different trajectories (one AC corresponding to each) were used to compute the various ACs. Note that the time lag units are different on each *x* axis.

unimportant and a first order Markovian SDE generated the observed noisy time series $\{z_{i\times n}\}_{i=0}^{N\times n}$, then the random force would need to have *all* of the statistical properties of a Brownian motion, namely, iid mean zero increments (not just uncorrelated increments). In this study, we know *a priori* that the z_t's are generated by a dynamical system where many degrees of freedom are not observed. The force ACs all die off relatively quickly and also give rise to very similar ACs, suggesting that the temporal correlation in the force is similar in each case analyzed. However, the channel has other unresolved slow degrees of freedom. Note how in Figure 1 the mean level changes after after 10−50 ps. The ion channel is flexible;[17] undulations of the protein and interactions of the tracked ion with the water chain and other ions in this narrow channel give rise to a more complex noise source. Artifacts of the non-Markovian (in 1-D) slow scale motion are reflected in the position autocorrelation. The three trajectories analyzed gave similar force ACs but very different ACs for the position. The shape of the position AC explains why a simple diffusive SDE is statistically rejected even at fairly large Δt values. Note that subdiffusive processes, e.g., fractional Brownian motion,[42] have been intentionally not considered as surrogates. [The fine structure apparent in the representative trajectory in Figure 1 suggests that a mathematically tractable subdiffusive process would not likely be able pass trajectorywise a goodness of fit test making full use of the implied conditional density of the assumed subdiffusive model. This author prefers the use of physically interpretable Markovian SDEs in part because measurement noise and other relevant physical features can be readily accounted for and powerful hypothesis testing machinery exists in this setting; recall that these tools check both the conditional distribution shape and assumed temporal correlations. If one can develop reliable tests checking various assumptions behind a particular non-Markovian surrogate and/or can demonstrate that such approaches make new useful physical predictions in a given system, these models should certainly be considered as potential surrogates, but model class selection is not the focus of this article.]

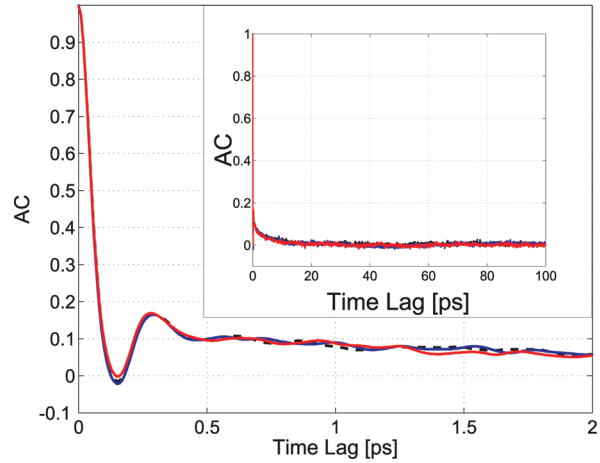


Figure 5. The position AC measured from three different 3 ns runs carried out in the presence of a constraining potential (see text). The inset shows the long time behavior, and the main portion of the figure zooms in only on early times.

One might inspect the position ACs in the unconstrained case and argue that the diffusion constant should be obtained using biased simulations where a harmonic guiding potential is employed in an attempt to constrain the dynamics close to a point of interest.[3,4] A physical motivation for this approach is to make the system's effective drift more closely resemble that associated with a 1-D harmonic potential and focus attention on the resulting effective diffusion.[4] Results using $k_{\text{harmonic}} = 40$ kcal/mol/Å^2 and adding the biasing potential $U_{\text{bias}}(z) = k_{\text{harmonic}}/2(z + 11\ \text{Å})^2$ to the MD evolution equations are reported to demonstrate that many of the previous complications observed before do not go away. The constrained results are qualitatively similar to those observed in the unconstrained case (see the Supporting Information); the one notable difference is in the position ACs. When three separate 3 ns constrained runs are analyzed, the resulting ACs obtained demonstrate "better" ergodic sampling in the sense that now the position ACs overlap substantially (see Figure 5). The global diffusion coefficient estimated using the method in ref 3 (using a 9 ns trajectory) was found to be 13.88 Å^2/ns and that obtained using the OU model was $\tilde{D} = 10.53 \pm 3.64$. Both diffusion coefficient values reported here are in close agreement with those reported in refs 4 and 17, computed using different computational methods also utilizing constrained simulations. Reference 4 used a memory kernel, whereas the values estimated here employed a

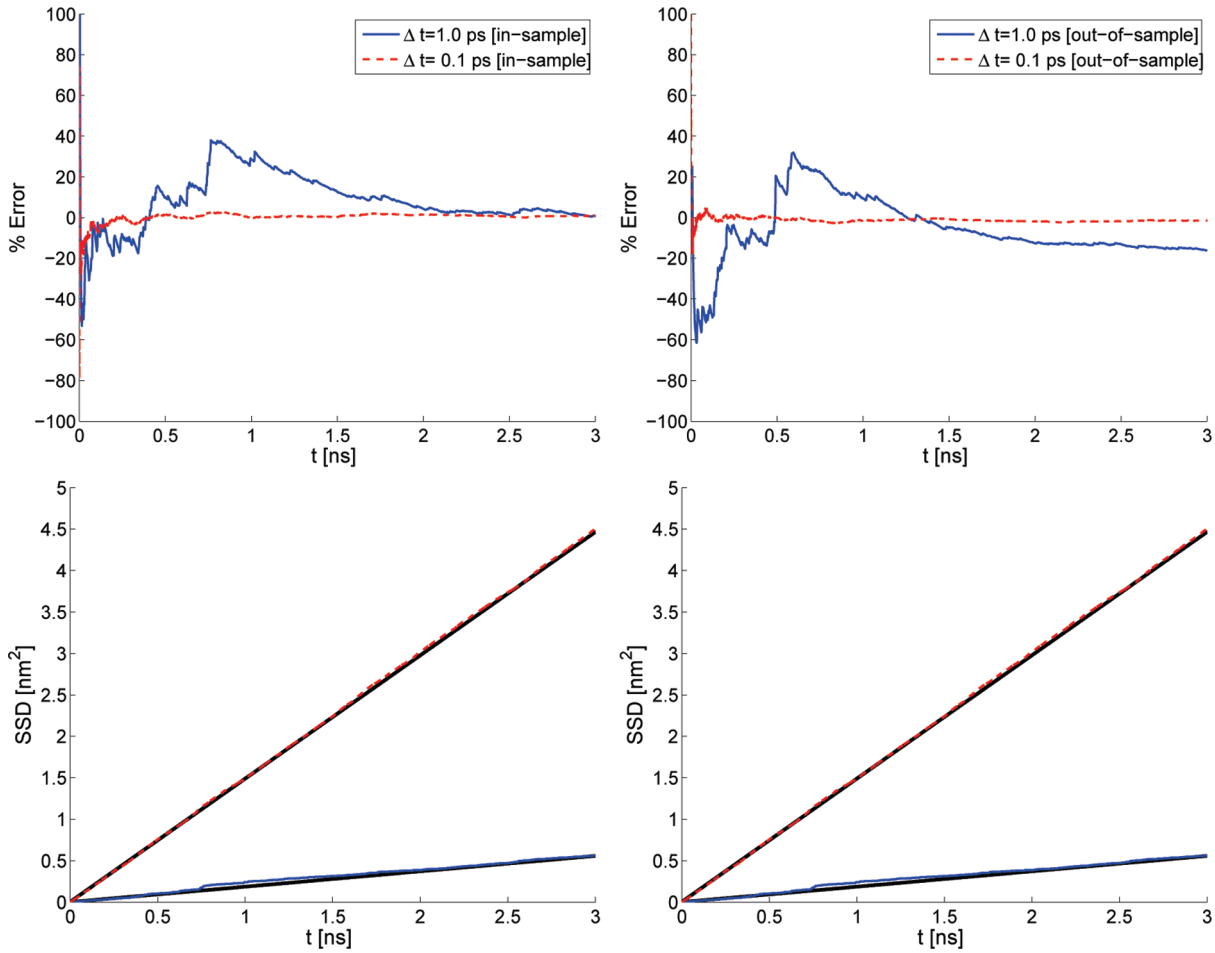


Figure 6. The percent error in the sum of squared displacements (top) and the raw observed and predicted sum of squared displacements (bottom). "In-sample" and "out-of-sample" data were analyzed (see text). The percent error was computed using $(\mathrm{SSD}^{\mathrm{observed}}(t) - \mathrm{SSD}^{\mathrm{predicted}}(t))/\mathrm{SSD}^{\mathrm{predicted}}(t) \times 100$.

Markovian SDE with downsampling (using $n = 100$) in one approach. The other computational method[3] reported in this paragraph utilized information extracted from the "converged" position AC.

Although these formal methods provide diffusion coefficients which are in agreement with one another, this does not imply that the computed quantity is consistent with observed trajectories. The Supporting Information shows that even for fairly large n or Δt even Markovian SDEs with the "correct" diffusion coefficient are still rejected using the M and Q tests in the constrained case. The agreement between formal methods for computing the effective diffusion coefficient does not imply predictive ability either. Furthermore, one is usually interested in computing the diffusion coefficient of the unconstrained system and using this result for predicting various physical quantities; so the artifacts observed in the unconstrained simulations should be dealt with in a surrogate model. That is, nontrivial fluctuations (or their damping rate) can make important contributions to forecasted events. One theme advocated in this article is that formal methods for computing the diffusion coefficient should be consistent with observation in some quantifiable sense and/or be able to make predictions (outside of the fitting criteria) of events occurring over time scales of interest. If neither criteria can be met, one should consider "non-traditional" approaches to computing the diffusion coefficient. For example, recent studies[43,44] demonstrated that the effective friction (which was related explicitly to the diffusion through the fluctuation dissipation theorem) of a coarse-grained model calibrated from observational data coming from high dimensional steered molecular dynamics simulations of a gramicidin channel could be used to predict mean first passage times under zero external force; the success of this approach relied on inferring the effective friction and force from a physically motivated fitting criterion different than formal procedures typically used in classic chemical physics computations.[44] Consequences of the fact that intermediate Δt's (in the 0.1−0.2 ps range) yield the best SDE proxy in the unconstrained case in the gA system studied here, in regards to both goodness of fit to the observed data and the predictive ability of statistics of the complex systems, for the regimes studied is the focus of the next two set of results. These findings also demonstrate that classic formal definitions of the diffusion coefficient should be reconsidered in low dimensional models or summaries of complex dynamical systems.

In order to demonstrate that simple SDE models calibrated at intermediate Δt values possess some predictive ability, Figure 6 plots the sum of squared displacements, $\mathrm{SSD}(t;n)$: $= 1/2\sum_{i=0}^{N_{\mathrm{sim}}(t)} (z_{(i+1)\times n} - z_{i\times n})^2$, coming from the MD simula-

tions and the straight line predicted by using the estimated $\tilde{D}$; $N_{sim}(t)$ corresponds to the number of time ordered observations generated by the MD simulation up to time t. Two regimes are studied, the intermediate regime providing the best fit, as judged by the Q and M statistics, and the larger $\Delta t \approx 1.0$ ps where the observed effective drift is low but the Q and M suggest something is askew. The complex unresolved slow-scale motion prevents a diffusion model from being statistically acceptable even for fairly large n (equivalently Δt). Said differently, the structured exploration of phase space (see the bottom panel of Figure 1) cannot be approximated by a process driven by Brownian motion. In order to demonstrate how these artifacts influence $\tilde{D}$'s ability to predict SSD, a 3 ns trajectory of simulation data was used to calibrate the parameters of the Ornstein−Uhlenbeck model. These calibration data are labeled as "in-sample", and another 3 ns of data (not used for parameter estimation) were labeled "out-of-sample". The Δt corresponding to the the lowest goodness of fit test statistics also provides the best predictive model. It is worth noting that the "best" prediction is judged in terms of percent error in the empirically in and out of sample SSD. In a parametric MLE estimate, drift and diffusion are both explicitly accounted for by the likelihood function (and also by the goodness of fit tests used), but in the SSD, the influence of the drift can adversely affect the SSD for larger Δt. However, the SSD plots suggest that these effects are not too dramatic. The rejection of the Ornstein−Uhlenbeck SDE calibrated using $n = 10$ with a moderately small sample size ($N = 100$) may be due to subjecting the model to an overly stringent hypothesis test; i.e., the errors in the SSD may be acceptable for a practical approximation in the physical sciences. However, it is nonetheless useful to know that the errors observed when using a diffusion approximation to describe increments of a more complex process are systematic and not simply sampling errors. Admittedly, predicting the SSD associated with the Δt yielding the best classic diffusion approximation may not be of interest in chemical applications per se. However, knowledge of the time scale where random forces can be approximated by a diffusion type processes has proven useful in making predictions relevant to nonequilibrium potential of mean force computations.[17]

It should be explicitly pointed out that the statistician's tenet of "thou shalt not waste data" was adhered to; i.e., even though subsampling occurred, each observation was eventually used for parameter estimation, e.g., see ref 23. In Figure 6, the slope of the line was predicted using the population average parameter observed using every observation in a single trajectory. The physical intuition behind using the population average implicitly assumes ergodic sampling (i.e., time averages are close to ensemble averages[16,45]). However, given that larger n eventually results in a rejected model, it would be interesting to see if there is enough statistical evidence to suggest that the estimated $\tilde{D}$ depends significantly on the full set of initial conditions for intermediate n. In in attempt to quantify this effect, sometimes referred to as "dynamic disorder",[45] the mixed effect model given in eq 4 was fit to the inferred local diffusion coefficient data. More specifically, an attempt was made to quantify if the variability induced by different initial conditions (drawn from a Boltzmann distribution) can be detected in the presence of sampling uncertainty. $N_{IC} = 20$ common position initial conditions were taken from the equilibrated initial condition, and the position coordinates were recorded every 100 ps in unconstrained simulations. From these multiple IC position files, $N_{Rep} = 10$ different random number streams and velocities' ICs were used to generate N_{Rep} short MD trajectories (i.e., a total of $N_{Rep} \times N_{IC} = 200$ trajectories of size $N = 100$ were analyzed). The subsampling parameter used here was $n = 10$ (corresponding to the "best" models as determined by the goodness of fit analysis presented earlier), and observations were recorded for each IC. The previous set of results demonstrated that quantities depending only on intermediate spacing between adjacent time series entries (recall that this spacing was quantified by $\Delta t = n\delta t$) had predictive ability. For these sampling parameters, there was sufficient evidence indicating the statistical significance of the random effect. The p value obtained when testing a mixed effect model versus a pure fixed effect reference model ($b_i^{\tilde{D}}$ was forced to be zero) was 0.043 (suggesting that the fixed effect model was suspect). The AIC and BIC also suggested that the random effect in the mixed effect model was statistically significant. In terms of statistical mechanics, this translates into the statement that variation induced by the different ICs is statistically significant; the distribution observed in the estimated local diffusion coefficients cannot be attributed to sampling uncertainty alone. A more intuitive demonstration is presented in the box plots of Figure 7 where $N_{IC} = 20$ different box plots are displayed. This lack of ergodic sampling might cause one to claim that this is "the sign of a bad reaction coordinate." However, it is important to attempt to make full use of time series information available; e.g., in single-molecule experiments, the observables available will likely be "imperfect" reaction coordinates.

Only fairly simple Ornstein−Uhlenbeck models were considered in the "practically stationary" regime in this article. This regime was studied mainly to facilitate the statistical analysis and demonstrate that nontrivial features can be detected even in this regime. The transition density, MLE, and associated limiting distribution can be computed in this case for the OU model without having to appeal to additional numerical approximations.[18] However, the tools presented can handle the addition of other features, such as time dependent external forces[13,15,17,29,33,34] and/or position dependent diffusion in a continuous time nonlinear SDE model.[12,16] The basic findings reported here come through even on SDE models with additional features. For example, box plots obtained by estimating the parameters of a continuous (time and state space) SDE taking position dependent "overdamped" noise into account[12] are shown to demonstrate that the mixed model analysis did not contain artifacts of missing position dependence in the local diffusion coefficient (the p value in the corresponding mixed effects analysis was 0.0033, and both the AIC and BIC favored the random effects model).

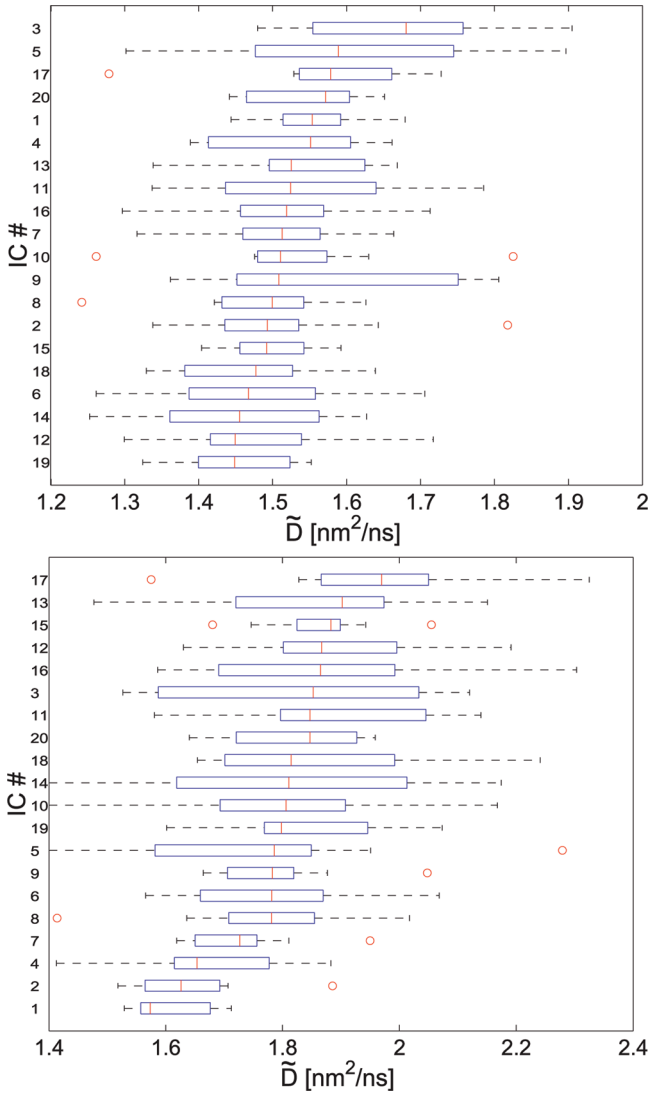


Figure 7. Box plots of $\tilde{D}$. Each box plot corresponds to a different initial condition drawn from an equilibrated MD simulation. The top plot corresponds to the Ornstein Uhlenbeck estimate and the bottom to a position dependent "overdamped" SDE (see text for additional details).

4. Conclusions and Outlook

Time domain maximum likelihood (transition density based) inference methods were demonstrated to be useful for fitting and assessing the statistical validity of a 1-D SDE model approximating the dynamics associated with gramicidin A simulations. Unresolved degrees of freedom were shown to be detectable on a "fast time scale" (time series observations were uniformly separated by time intervals < 0.1ps) where force correlations were found to be statistically significant[4] and also at "slow time scales" studied (time between adjacent observations ranging from 0.8 to 60 ps). At intermediate time scales, a collection of *local* SDEs was shown (1) to be a better statistical summary of the data as measured by goodness of fit tests which made use of the entire assumed conditional distribution and time correlations, (2) to contain enough statistical evidence to indicate that nonergodic sampling was occurring (the mixed effects model approach was also shown useful in quantifying the degree of the initial condition dependence), and (3) to have predictive capability for out-of-sample data; it should be noted that the predicted quantity was not used as a fitting criterion.

It is not too surprising that solely monitoring the axial location of a tagged ion is problematic in defining a 1-D diffusion coefficient in an ion channel where channel undulations, nontrivial solvation effects, and other unresolved factors modulate the dynamics.[2,17,20,45] These factors can significantly complicate estimation of the system's position autocorrelation (as shown here and in ref 16) and prediction of more complex long-term events (such as mean first passage times). However, assessing the validity of various simplifying assumptions, such as the suitability of an assumed Markovian model on a specified time scale[35,36] given time ordered observations, will assist in better understanding/summarizing the rich information coming from all-atom computer simulations and single-molecule experiments.[12,14,28] The utility of frequentist inferential methods employing transition density information in analyzing the effective diffusive noise as opposed to fitting an autocorrelation function or only focusing on some other low order stationary moments was also demonstrated (aspects of this issue are discussed more extensively in ref 32). In regard to some traditional chemical physics computations, such as mean first passage time computations,[3,6,46] often computations of both the effective diffusion coefficient and free energy differences are required. In such cases, it is possible that systematic errors in both the classic diffusion coefficient and free energy estimates can cancel to provide the same mean first passage time ("rate") prediction. It is useful to have reliable criteria for checking various model assumptions with hypothesis testing machinery.[28] The methods presented here can be used to determine if the implicit assumptions behind a given coarse system description[3,6,46] are appropriate given observational data. Careful statistical analysis of the diffusion coefficient can potentially help in assessing the accuracy of estimated free energy or PMF differences (indirectly) if one is only given experimental flux measurements. Such analyses can potentially identify factors that may be confounded in traditional mean first passage time analyses. If formal definitions for the diffusion coefficient are not consistent with data and/or unable to make useful predictions, one should consider alternate approaches for quantifying "thermal noise", as demonstrated here and in refs 44 and 45.

With the advent of single-molecule experiments and ever increasing MD simulation power, it is important to consider physically interpretable data summaries that possess predictive ability if we hope to fully utilize the wealth of information coming from these new data sources. The need for models that have testable criteria which can be applied to both experimental and simulation time series poses new and exciting challenges in describing biological systems.[12,14,44,47] For simplicity, the focus here was on a roughly stationary signal [that is, the moments of the time series were roughly time independent[4]] approximated by an SDE with a constant diffusion coefficient, variants of the MLE type of approach are applicable to nonergodic cases where time dependent external forces

are added into the system and the local diffusion coefficient depends on the value of the order parameter.[33,34] Similar approaches have been demonstrated to be useful in understanding experimental data where measurement noise (on top of thermal noise) is also present.[13−15] The magnitude of the thermal and measurement noise can both be fit from observational data (these quantities do not need to be guessed or assumed *a priori*), and the goodness of fit tests can be used to determine if a proposed model is appropriate given the data. For example, one can explicitly test if thermal noise dominates measurement noise without requiring an implicit stationarity assumption.[14] Fourier transform based methods, popular in statistical physics, often require such stationary assumptions, but these can be hard to satisfy in single-molecule data.[13,15] With time domain likelihood based approaches, dynamic signatures of unresolved degrees of freedom have been suggested by the estimated position dependent diffusion coefficient in studies analyzing experimental data where time-dependent forces are added, e.g., see refs 13 and 14. The type of data summary presented here, where the entire distribution implied by an assumed surrogate model is used to assess the fit and a mixed effects model is used to respect the variability induced by a lack of ergodic sampling shows promise in understanding complex data sets arising from future simulations and experiments. Attempts were made to avoid appealing to "memory kernels"[4] or long memory processes in order to facilitate the physical interpretation of the surrogate SDEs and utilize information that is experimentally accessible (e.g., force or position). Regardless of the type of surrogate model used (i.e., one with or without memory), mixed effect modeling techniques show promise as tools for quantitatively summarizing the dynamics observed when unresolved degrees of freedom are believed to be important but not directly measurable.[34]

Acknowledgment. The author obtained partial computational support from the Rice Computational Research Cluster funded by NSF under Grant CNS-0421109 and a partnership between Rice University, AMD, and Cray.

Supporting Information Available: The Ornstein−Uhlenbeck model, goodness of fit tests, rough guidelines to more general SDE modeling, and mixed effects models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Roux, B.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 4856–4868.

(2) Burykin, A.; Kato, M.; Warshel, A. *Proteins* **2003**, *52*, 412–426.

(3) Hummer, G. *New J. Phys.* **2005**, *7*, 34.

(4) Mamonov, A.; Kurnikova, M.; Coalson, R. *Biophys. Chem.* **2006**, *124*, 268–278.

(5) Forney, M. W.; Janosi, L.; Kosztin, I. *Phys. Rev. E* **2008**, *78*, 051913.

(6) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–13846.

(7) Moffitt, J.; Chemla, Y.; Smith, S.; Bustamante, C. *Annu. Rev. Biochem.* **2008**, *77*, 19.119.4.

(8) Nollmann, M.; Stone, M. D.; Bryant, Z.; Gore, J.; Crisona, N. J.; Hong, S. C.; Mitelheiser, S.; Maxwell, A.; Bustamante, C.; Cozzarelli, N. R. *Nat. Struct. Mol. Biol.* **2007**, *14*, 264–271.

(9) Walther, K.; Gräter, F.; Dougan, L.; Badilla, C.; Berne, B.; Fernandez, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7916–21.

(10) Greenleaf, W.; Frieda, K.; Foster, D.; Woodside, M.; Block, S. *Science* **2008**, *319*, 630–633.

(11) Hodges, C.; Bintu, L.; Lubkowska, L.; Kashlev, M.; Bustamante, C. *Science* **2009**, *325*, 626–628.

(12) Calderon, C.; Arora, K. *J. Chem. Theory Comput.* **2009**, *5*, 47.

(13) Calderon, C.; Harris, N.; Kiang, C.; Cox, D. *J. Phys. Chem. B* **2009**, *113*, 138.

(14) Calderon, C.; Chen, W.; Harris, N.; Lin, K.; Kiang, C. *J. Phys.: Condens. Matter* **2009**, *21*, 034114.

(15) Calderon, C.; Harris, N.; Kiang, C.; Cox, D. *J. Mol. Recognit.* **2009**, *22*, 356.

(16) Calderon, C. P. *Phys. Rev. E* **2009**, *80*, 061118.

(17) Calderon, C.; Janosi, L.; Kosztin, I. *J. Chem. Phys.* **2009**, *130*, 144908.

(18) Tang, C.; Chen, S. *J. Econometrics* **2009**, *149*, 65–81.

(19) Allen, T. W.; Bastug, T.; Kuyucak, S.; Chung, S. H. *Biophys. J.* **2003**, *84*, 2159–2168.

(20) Miloshevsky, G. V.; Jordan, P. C. *Biophys. J.* **2004**, *86*, 92–104.

(21) Allen, T. W.; Andersen, O. S.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117–122.

(22) Braun-Sand, S.; Burykin, A.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 583–592.

(23) Zhang, L.; Mykland, P.; Ait-Sahalia, Y. *J. Am. Stat. Assoc.* **2005**, *100*, 1394–1411.

(24) Pavliotis, G. A.; Stuart, A. M. *J. Stat. Phys.* **2007**, *127*, 741–781.

(25) Calderon, C. *Multiscale Model. Simul.* **2007**, *6*, 656–687.

(26) Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, D. *nlme: Linear and Nonlinear Mixed Effects Models*, R package version 3.1−96; R Core team: 2009.

(27) Ruppert, D.; Wand, M.; Carroll, R. *Semiparametric Regression*; Cambridge University Press: New York, 2003; pp 91−110.

(28) Calderon, C. P. *J. Phys. Chem. B* **2010**, *114*, 3242–3253.

(29) Calderon, C.; Chelli, R. *J. Chem. Phys.* **2008**, *128*, 145103.

(30) Burykin, A.; Kato, M.; Warshel, A. *Proteins* **2003**, *52*, 412–426.

(31) Smith, G. R.; Sansom, M. S. *Biophys. Chem.* **1999**, *79*, 129–151.

(32) Pokern, Y.; Stuart, A.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *8*, 69–95.

(33) Calderon, C. *J. Chem. Phys.* **2007**, *126*, 084106.

(34) Calderon, C.; Martinez, J.; Carroll, R.; Sorensen, D. *Multiscale Model. Simul.* **2010**, *8*, 1562–1580.

(35) Hong, Y.; Li, H. *Rev. Fin. Studies* **2005**, *18*, 37–84.

(36) At-Sahalia, Y.; Fan, J.; Jiang, J. *Annals of Statistics* **2010**, *38*, 3129−3163.

(37) Protter, P. E. *Stochastic Integration and Differential Equations*, 2nd ed.; Springer: New York, 2003; pp 12−84.

(38) Chen, S.; Gao, J.; Tang, C. *Ann. Stat.* **2008**, *36*, 167–198.

(39) Johnson, V. E. *Ann. Stat.* **2004**, *32*, 2361.

(40) Ait-Sahalia, Y.; Fan, J.; Peng, H. *JASA* **2009**, *104*, 1102–1116.

(41) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(42) Kou, S.; Xie, X. *Phys. Rev. Lett.* **2004**, *93*, 180603.

(43) Kamerlin, S.; Vicatos, S.; Dryga, A.; Warshel, A. *Annu. Rev. Phys. Chem.* **2011**; doi: 10.1146/annurev-physchem-032210-103335.

(44) Dryga, A.; Warshel, A. *J. Phys. Chem. B* **2010**, *114*, 12720−12728.

(45) Kuo, T. L.; Garcia-Manyes, S.; Li, J.; Barel, I.; Lu, H.; Berne, B. J.; Urbakh, M.; Klafter, J.; Fernandez, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 11336–11340.

(46) Pisliakov, A. V.; Cao, J.; Kamerlin, S. C.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, 17359–17364.

(47) Stock, G.; Ghosh, K.; Dill, K. *J. Chem. Phys.* **2008**, *128*, 194102.

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

CT1004966

JCTC Journal of Chemical Theory and Computation

Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals—Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions

Lars Goerigk[†,‡] and Stefan Grimme*[,†]

Theoretische Organische Chemie, Organisch—Chemisches Institut der Universität Münster, Corrensstraβe 40, and NRW Graduate School of Chemistry, Wilhelm-Klemm-Straβe 10, D-48149 Münster, Germany

Received August 18, 2010

Abstract: We present an extended and improved version of our recently published database for general main group thermochemistry, kinetics, and noncovalent interactions [*J. Chem. Theory Comput.* **2010**, *6*, 107], which is dubbed GMTKN30. Furthermore, we suggest and investigate two new double-hybrid-meta-GGA density functionals called PTPSS-D3 and PWPB95-D3. PTPSS-D3 is based on reparameterized TPSS exchange and correlation contributions; PWPB95-D3 contains reparameterized PW exchange and B95 parts. Both functionals contain fixed amounts of 50% Fock-exchange. Furthermore, they include a spin-opposite scaled perturbative contribution and are combined with our latest atom-pairwise London-dispersion correction [*J. Chem. Phys.* **2010**, *132*, 154104]. When evaluated with the help of the Laplace transformation algorithm, both methods scale as N^4 with system size. The functionals are compared with the double hybrids B2PLYP-D3, B2GPPLYP-D3, DSD-BLYP-D3, and XYG3 for GMTKN30 with a quadruple-ζ basis set. PWPB95-D3 and DSD-BLYP-D3 are the best functionals in our study and turned out to be more robust than B2PLYP-D3 and XYG3. Furthermore, PWPB95-D3 is the least basis set dependent and the best functional at the triple-ζ level. For the example of transition metal carbonyls, it is shown that, mainly due to the lower amount of Fock-exchange, PWPB95-D3 and PTPSS-D3 are better applicable than the other double hybrids. Finally, we discuss in some detail the XYG3 functional [*Proc. Nat. Acad. Sci. U.S.A.* **2009**, *106*, 4963], which makes use of B3LYP orbitals and electron densities. We show that it is basically a highly nonlocal variant of B2PLYP and that its partially good performance is mainly due to a larger effective amount of perturbative correlation compared to other double hybrids. We finally recommend the PWPB95-D3 functional in general chemistry applications.

1. Introduction

Kohn−Sham density functional theory (KS-DFT)[1−5] has become the "work-horse" of modern quantum chemistry. It represents a good compromise between computational effort and accuracy. Whenever costly wave function based methods are not applicable to a certain kind of chemical problem, DFT provides a valuable alternative. However, the huge number of developed density functionals (DFs) to date shows that current approximate DFT still suffers from several flaws and that the quest for finding a functional, which comes close to the "true one", is still ongoing. In this context, we want to particularly focus on the fact that not every DF is equally applicable to every problem (see, e.g., ref 6 for further information). This makes choosing the right functional for the right problem a tough task, even for experienced

* Corresponding author phone: (+49)-251-8333241, e-mail: grimmes@uni-muenster.de.

† Universität Münster.

‡ NRW Graduate School of Chemistry.

10.1021/ct100466k
Published on Web 12/23/2010

researchers in this field. Therefore, we think that the development of highly accurate and concomitantly robust, i.e., broadly applicable, DFs is very desirable.

In 2006, an important step toward this aim was the development of the B2PLYP double-hybrid density functional (DHDF),[7] which has its roots in the Görling−Levy Kohn−Sham perturbation theory.[8,9] It combines a standard hybrid-GGA DFT calculation with a second-order perturbative treatment based on KS orbitals, thus introducing nonlocal correlation effects or, in other words, information about virtual KS orbitals. For related precursors of this method, that mix wave function (WF) and DFT parts, see refs 10−12. Soon after, several variants of the double-hybrid idea were published.[13−23] The superior performance of double hybrids compared to common DFs was proven in many applications.[24−40]

The most recent study, showing the accuracy and robustness of B2PLYP, was at the same time the most thorough one. It was based on the so-called GMTKN24 database, which is a collection of 24 previously published or newly developed benchmark sets for general main group thermochemistry, kinetics, and noncovalent interactions.[35] It covers atomization energies, electron affinities, ionization potentials, proton affinities, self-interaction error (SIE) related problems, barrier heights, various reaction energies, particularly difficult cases for DFT methods, relative energies between conformers, and inter- and intramolecular noncovalent interactions. We pointed out that the range of properties covered by the GMTKN24 data set outperforms, to the best of our knowledge, all other combinations of databases that had been previously proposed. The GMTKN24 database's composition reflects many years of experience in benchmarking and in the application of DFT methods to "real-life" chemical problems. Our further positive experience with GMTKN24, since its publication, encouraged and confirmed our first impression that it is highly representative for chemistry (excluding transition metal compounds). Any quantum chemical method, that performs well for the entire database, can be really regarded as an accurate, robust, and useful method.

Since the publication of GMTKN24, we regarded six newly published benchmark sets as useful for giving further insight into a functional's applicability and performance.[41−43] Furthermore, recent developments and findings made it necessary to modify three of the original subsets.[44,45] Herein, we present the extended and modified version of GMTKN24, which is from now on called GMTKN30 and is recommended as a replacement.

With the help of GMTKN30, we want to re-evaluate the B2PLYP functional and want to compare the results with the recently published double hybrids B2GPPLYP,[15] DSD-BLYP,[20] and XYG3.[21] Furthermore, we also present two new DHDFs, called PTPSS and PWPB95, which are not based on hybrid-GGA but on hybrid-meta-GGA ingredients. For these two DHDFs, the perturbative treatment is carried out within a spin-opposite scaled (SOS) scheme.[46,47] When combined with a Laplace transformation algorithm,[48] this reduces the formal computational cost from N^5, with N being the system size, to N^4, which is then formally the same as the scaling of common hybrid DFs. Similar ideas have recently been proposed by Scuseria et al. in the framework of a truncated random phase approximation ansatz combined with long-range corrected DFT parts (LC-ωLDA+JMP2).[49]

This manuscript is structured as follows. First, the extended GMTKN30 database is presented. Second, the background of double-hybrid density functional theory will be discussed, with an emphasis on the XYG3 variant, and the new PTPSS and PWPB95 methods. Together with B2PLYP, B2GPPLYP, and DSD-BLYP, these methods are then benchmarked against GMTKN30. Moreover, we want to give an impression of the functionals' performance for transition metal chemistry with the example of carbonyl dissociation reactions. Furthermore, we will suggest a new scheme with which to determine the s_6 scale parameter of the DFT-D3 correction for DHDFs. We will also study basis set effects, address critical points recently raised against the DHDF formalism,[45,21,22,50] and shed some light on the XYG3 approach.

2. The GMTKN30 Data Set

The recently presented GMTKN24 database for general main group thermochemistry, kinetics, and noncovalent interactions covers a large variety of 24 different, chemically relevant subsets.[35] Here, we present an extended version with six new and three modified subsets. This extended database is called GMTKN30. In Table 1, short descriptions for each part of the GMTKN30 database are given, including the number of entries, a specification of the reference values, and the relevant citations. Note that none of the reference data include zero point vibrational energies (ZPVEs) or thermal (enthalpic) corrections. The type and source of reference data are given separately for each subset. In total, the GMTKN30 database comprises 1218 single point calculations and 841 data points (relative energy values). In Figure 1, an overview of the six additional subsets is given. For each set, the, on average, easiest and most difficult reactions (for GGA functionals) are shown. The reference values and the optimized coordinates of all systems are available for download from our Web site.[51] A detailed description of the original GMTKN24 subsets is given in ref 35. In the following, only the changes and additions to the original database are described.

2.1. The Modified NBPRC Subset. For the GMTKN24 database, a completely new benchmark set, called NBRC, was introduced.[35] It contained six oligomerization and dihydrogen fragmentation reactions of ammonia/borane systems. The reason for creating such a set was previous evidence of poor performance for similar reactions for some popular density functionals like B3LYP (see, e.g., ref 52). Recently, we investigated the mechanism of H_2 activation by frustrated Lewis pairs (FLPs).[44] In order to validate the theoretical methods used in that study, a small benchmark set for H_2 activation by three FLP-like model systems was developed:

$$PR_3^1 + BR_3^2 \rightarrow R_3^1P - BR_3^2 \qquad (A)$$

Table 1. Description of the Subsets within the GMTKN30 Database (New or Modified Subsets Are Emphasized in Italics)

set	description	#	av. $\|\Delta E\|$[a]	ref method	reference
MB08-165	decomposition energies of artificial molecules	165	117.2	est. CCSD(T)/CBS	*b*
W4-08	atomization energies of small molecules	99	237.5	W4	*c*
W4-08woMR	W4-08 without multireference cases	83	261.5	W4	*c*
G21IP	adiabatic ionization potentials	36	250.8	exp.	*d*
G21EA	adiabatic electron affinities	25	33.6	exp.	*d*
PA	adiabatic proton affinities	12	174.9	est. CCSD(T)/CBS and W1	*e*, *f*
SIE11	self-interaction error related problems	11	34.0	est. CCSD(T)/CBS	*g*
BHPERI	barrier heights of pericyclic reactions	26	19.4	W1 and CBS-QB3	*c*, *h*, *i*, *j*, *k*
BH76	barrier heights of hydrogen transfer, heavy atom transfer, nucleophilic substitution, unimolecular, and association reactions	76	18.5	W1 and theor. est.	*l*, *m*
BH76RC	reaction energies of the BH76 set	30	21.5	W1 and theor. est.	*l*, *m*
RSE43	radical stabilization energies	43	7.5	est. CCSD(T)/CBS	*n*
O3ADD6	reaction energies, barrier heights, association energies for addition of O_3 to C_2H_4 and C_2H_2	6	22.7	est. CCSD(T)/CBS	*o*
G2RC	reaction energies of selected G2/97 systems	25	50.6	exp.	*p*
AL2X	dimerization energies of AlX_3 compounds	7	33.9	exp.	*q*
NBPRC	oligomerizations and H_2 fragmentations of NH_3/BH_3 systems; H2 activation reactions with PH_3/BH_3 systems	12	27.3	est. CCSD(T)/CBS	*g*, *r*
ISO34	isomerization energies of small and medium-sized organic molecules	34	14.3	exp.	*s*
ISOL22	isomerization energies of large organic molecules	22	18.3	SCS-MP3/CBS	*t*
DC9	nine difficult cases for DFT	9	35.7	theor. and exp.	*g*, *j*, *u*, *v*, *w*, *x*, *y*, *z*
DARC	reaction energies of Diels−Alder reactions	14	32.2	est. CCSDT/CBS	*q*
ALK6	fragmentation and dissociation reactions of alkaline and alkaline−cation−benzene complexes	6	44.6	est. CCSD(T)/CBS	*aa*
BSR36	bond separation reactions of saturated hydrocarbons	36	16.7	est. CCSD(T)/CBS	*bb*
IDISP	intramolecular dispersion interactions	6	13.5	theor. and exp.	*s*, *cc*, *dd*, this work
WATER27	binding energies of water, $H^+(H_2O)_n$ and $OH^-(H_2O)_n$ clusters	27	82.0	est. CCSD(T)/CBS; MP2/CBS	*ee*
S22	binding energies of noncovalently bound dimers	22	7.3	est. CCSD(T)/CBS	*ff*, *gg*
ADIM6	interaction energies of *n*-alkane dimers	6	3.3	est. CCSD(T)/CBS	*aa*
RG6	interaction energies of rare gas dimers	6	0.46	exp.	*aa*, *hh*, *ii*, *jj*, *kk*
HEAVY28	noncovalent interaction energies between heavy element hydrides	28	1.3	est. CCSD(T)/CBS	*aa*
PCONF	relative energies of phenylalanyl−glycyl−glycine tripeptide conformers	10	1.5	est. CCSD(T)/CBS	*ll*
ACONF	relative energies of alkane conformers	15	1.8	W1h-val	*mm*
SCONF	relative energies of sugar conformers	17	4.9	est. CCSD(T)/CBS	*g*, *nn*
CYCONF	relative energies of cysteine conformers	10	2.1	est. CCSD(T)/CBS	*oo*

[a] Averaged absolute energies in kcal mol^{-1}, excluding ZPVEs. [b] Ref 31. [c] Ref 15. [d] Ref 116. [e] Ref 117. [f] Ref 118. [g] Ref 34. [h] Ref 119. [i] Ref 120. [j] Ref 25. [k] Ref 121. [l] Ref 122. [m] Ref 123. [n] Ref 124. [o] Ref 125. [p] Ref 126. [q] Ref 127. [r] Ref 43. [s] Ref 60. [t] Ref 40. [u] Ref 128. [v] Ref 129. [w] Ref 130. [x] Ref 131. [y] Ref 132. [z] Ref 7. [aa] Ref 42. [bb] Ref 41. [cc] Ref 23. [dd] Ref 55. [ee] Ref 133. [ff] Ref 58. [gg] Ref 44. [hh] Ref 66. [ii] Ref 67. [jj] Ref 68. [kk] Ref 69. [ll] Ref 134. [mm] Ref 135. [nn] Ref 136. [oo] Ref 137.

$$R^1_3P - BR^2_3 + H_2 \rightarrow [PR^1_3H]^+[BR^2_3H]^- \qquad (B)$$

Reaction A describes the formation of the Lewis pairs, whereas B is their reaction with H_2. We considered three reactions A and B with $R^1 = H/R^2 = H$, $R^1 = CH_3/R^2 = F$, and $R^1 = CH_3/R^2 = Cl$. Geometries for these model reactions were obtained at the B3LYP-D/TZVP level of theory. Estimated CCSD(T)/CBS reference values for these reactions were obtained as proposed by Jurecka and Hobza.[53] MP2/CBS[54] values (based on cc-pVTZ and cc-pVQZ results) were corrected by the difference of CCSD(T)/cc-pVTZ and MP2/cc-pVTZ correlation energies. We added these reactions to the original set and dubbed the new set NBPRC. In order to calculate all reaction energies of NBPRC, 21 single point calculations have to be carried out. The energy range is from −48.3 to +40.4 kcal/mol. The average absolute reaction energy is 27.3 kcal/mol.

2.2. Changes to the IDISP Subset. The original IDISP subset for intramolecular London-dispersion effects of six large organic systems involves 13 single point calculations and has an average relative energy of 14.1 kcal/mol.[35] In three cases, we felt it necessary to recalculate the reference values. The original reference value of 9.4 kcal/mol for the isomerization of *n*-undecane to 2,2,3,3,4,4-hexamethylpentane was based on the SCS-MP2/cQZV3P//MP2/TZVP level of theory.[55] We found significant differences between MP2/CBS (3.9 kcal/mol) and SCS-MP2/CBS (10.0 kcal/mol) treatments (based on aug-cc-pVTZ → aug-cc-pVQZ extrapolations). Thus, we decided to base the new value on MP2.5; i.e., an MP2/CBS energy is combined with one-half of the third-order contribution of MP3/aug-cc-pVDZ (carried out with the group's own program RICC[56]). MP2.5 had been shown to yield accurate results that are in some cases even comparable to a CCSD(T)/CBS treatment.[57] The new reference value is 8.2 kcal/mol. Furthermore, the reference values for the folding of the $C_{14}H_{30}$ and $C_{22}H_{46}$ hydrocarbons were recalculated. Originally, the values of −2.2 and +3.6 kcal/mol were based on MP2/aug-cc-pVTZ//BLYP-D/TZV(p,d) calculations.[24] However, according to MP2/CBS results, convergence to about 0.5 kcal/mol accuracy is only obtained after aug-cc-pVTZ → aug-cc-pVQZ extrapolations. The new values are −3.1 and +0.4 kcal/mol, respectively. Because of the new reference values, the average relative energy for

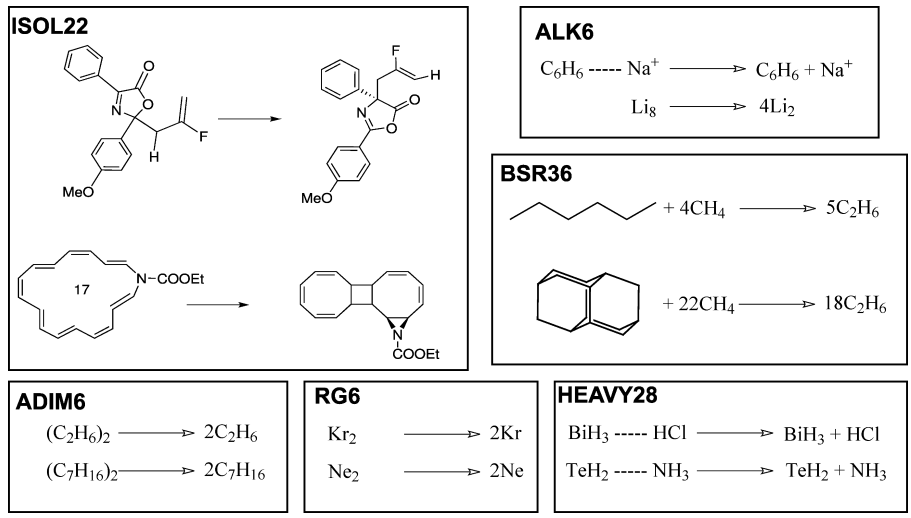


Figure 1. The six new subsets of the GMTKN30 database. For each set, the, on average, easiest (top) and most difficult (bottom in each box) reactions (for GGAs) are shown.

the complete subset changes to 13.5 kcal/mol. The energy range is from −58.5 to +8.2 kcal/mol.

2.3. New Reference Values for the S22 Set. Hobza and co-workers derived the reference values for the 22 interaction energies of noncovalently bound complexes (S22 set) within an estimated CCSD(T)/CBS scheme.[58] However, as Sherrill and co-workers argued, both MP2/CBS values and the differences between CCSD(T) and MP2 correlation energies were based on various basis sets for different systems and, thus, are not consistent throughout the set. Therefore, they recently estimated new CCSD(T)/CBS data.[45] Shortly after that publication, Podeszwa et al. also proposed new reference values.[59] We considered both proposals carefully and found the two sets of new reference values to be almost identical. We will use the energies published by Sherrill and co-workers. With these revised values, the S22 set has an average absolute interaction energy of 7.3 kcal/mol. The energy range is from 0.5 to 20.7 kcal/mol.

2.4. The ISOL22 Subset. Very recently, Huenerbein et al. published a new benchmark set containing 24 isomerization reactions (ISOL[41]) of large molecules covering a wide range of different compounds, like, e.g., a sugar, a steroid, an organic dye, hydrocarbons, and large molecules containing heteroatoms. In contrast to the popular ISO34 set,[60] which is also a part of GMTKN24 and GMTKN30, the large size of the molecules casts an additional light on effects that are important in "real life" organic chemistry. These are, in particular, intramolecular London-dispersion effects. Furthermore, charged systems are also considered. Reference values are based on the SCS-MP3/CBS//B97-D/TZVP levels of theory. For the present study, we excluded reactions **1** and **4** (see ref 41 for more details) as treating them is very time-consuming and not feasible in an extensive benchmark study. Thus, the subset presented herein contains only 22 reactions (44 single point calculations) and is called ISOL22. The energy range is from 0.5 to 38.1 kcal/mol. The average reaction energy is 18.3 kcal/mol.

2.5. ALK6. For the development of the new London-dispersion correction termed DFT-D3, Grimme et al. introduced the so-called ALK6 benchmark set that includes three decomposition reactions of alkaline metal complexes M_8 (M = Li, Na, K) into their dimers and three dissociation reactions of alkaline-cation−benzene complexes M^+−Bz.[43] Reference values are based on estimated CCSD(T)/CBS calculations. The complete set comprises 13 single point calculations, and its average reaction energy is 44.6 kcal/mol. The energy range is from 19.2 to 83.2 kcal/mol.

2.6. The BSR36 Subset. Recently, Steinmann et al.[61] carried out a dispersion corrected density functional study on 36 bond separation reactions (as introduced by Pople and co-workers[62,63] and also recently investigated by Wodrich et al.[64]). These are reactions of different saturated hydrocarbons [15 (partially branched) chains, five cages, and 16 rings] with methane to yield ethane. As reference values, experimental heats of formation were taken. However, Krieg and Grimme revealed that usage of these reference values led to misleading interpretations regarding different density functionals.[42] They concluded that a theoretical reference is more appropriate for this test set and computed reaction energies on the estimated CCSD(T)/CBS//MP2/cc-pVTZ level of theory. The complete test set was dubbed BSR36. The set comprises 38 single point calculations and has an average reaction energy of 16.7 kcal/mol. The energy range is from 2.4 to 51.4 kcal/mol.

2.7. The ADIM6 Subset. Tsuzuki et al. published estimated CCSD(T)/CBS reference values for the interaction energies of *n*-alkane dimers ($n = 1-10$).[65] Grimme et al. took the systems with $n = 2-7$ for their study of the new London-dispersion correction and called this benchmark set ADIM6.[42] ADIM6 involves 12 single point calculations and has an average interaction energy of 3.3 kcal/mol. The energy range is from 1.3 to 5.6 kcal/mol.

2.8. The RG6 Subset. Grimme et al. used (partially theoretically corrected) experimental[66−69] dissociation energies of five homonuclear and one heteronuclear rare gas dimer for the development of DFT-D3.[43] This subset was denoted RG6. It involves 11 single point calculations and has an average dissociation energy of 0.46 kcal/mol. The energy range is from 0.08 to 0.79 kcal/mol.

2.9. The HEAVY28 Subset. The HEAVY28 benchmark set by Grimme et al. comprises 28 noncovalent interaction energies of different heavy element hydrides (e.g., including hydrides of Sb, Te, I, Pb, and Bi).[43] Reference values are based on estimated CCSD(T)/CBS calculations. Thirty-eight single point calculations are carried out for the evaluation of HEAVY28. The energy range is from 0.44 to 3.29 kcal/mol. Its average interaction energy is 1.31 kcal/mol.

2.10. Weighted Total Mean Absolute Deviation. In our recent study, we completed the analysis of the GMTKN24 database using an overall statistical evaluation. In the spirit of the work by Truhlar and co-workers (see, e.g., ref 70), we defined a weighted total mean absolute deviation (WTMAD) to combine all obtained mean absolute deviations (MADs) for each subset into one final number for a tested method. We also discussed that such a procedure can be defined in several ways and that there is no real right or wrong. After having tested several schemes, we found that the overall ranking of methods was not altered. In the scheme, which we finally presented (see eq 1), each of 24 MAD values was weighted by the number of entries (N_i) of each subset to take into account its size. Furthermore, each subset was weighted by an additional factor that was calculated as the ratio between the MADs of BLYP and B2PLYP-D [i.e., MAD(BLYP)/MAD(B2PLYP-D)] to take into account the difficulty of a certain subset.

$$\mathrm{WTMAD} = \frac{1}{3091.4} \times \sum_{i}^{30} N_i \times \frac{\mathrm{MAD}_i^{\mathrm{BLYP}}}{\mathrm{MAD}_i^{\mathrm{B2PLYP-D}}} \times \mathrm{MAD}_i \tag{1}$$

In order to be consistent, and although we will apply the new DFT-D3 correction in the present study, we decided to define the WTMAD for GMTKN30 in the same way, i.e., with the older version DFT-D. The only difference is that we introduced the additional constraint that the product of system size and scale factor of a certain set should not be larger than one-half of the corresponding value for MB08-165, i.e., 222.75. Therefore, it is guaranteed that smaller sets with a large scale factor enter not too strongly. The actual values for the weighting factors of all 30 subsets are given in the Supporting Information.

3. Double-Hybrid Density Functional Theory

Double-hybrid density functionals (DHDFs) are situated on the fifth rung in Perdew's scheme of "Jacob's ladder"[71] as they include virtual Kohn−Sham orbitals (here, we also want to acknowledge that the closely related term "doubly hybrid" originated from Truhlar and co-workers' multicoefficient methods[10,11]). Compared to hybrid-GGA functionals (fourth rung), where some part of the exchange functional is substituted by "exact" (HF) exchange, DHDFs additionally substitute some part of the correlation functional by mixing in a nonlocal perturbative correlation. This correlation part is basically obtained by a second-order Møller−Plesset (MP2) type treatment based on KS orbitals and eigenvalues. The first DHDF following this idea is the B2PLYP functional of Grimme,[7] which was soon followed by the mPW2PLYP functional of Schwabe and Grimme.[13] B2PLYP is nowadays a widely recognized functional, which, in combination with an empirical London-dispersion term (DFT-D[72]/DFT-D3[43]), resulted in being very accurate and robust in several ground- and excited-state studies.[24−40,43] This stimulated further works, and several modifications of B2PLYP were proposed in recent years. These are the B2KPLYP, B2TPLYP, and B2GPPLYP variants by Martin and co-workers, specifically designed as functionals working well for kinetics, thermochemistry, and general purpose applications.[14,15] The reparameterized B2π-PLYP functional of Sancho-García and Péréz-Jimenéz was developed to work particularly well in π-conjugated systems.[16] Head-Gordon and co-workers proposed a distance-dependent scaling of the perturbative correlation part and developed the variants B2P3LYP and B2OS3LYP.[17] The latter functional includes perturbative contributions of electron pairs with opposite spins only (in the spirit of the SOS-MP2[47] method). Radom and co-workers proposed a reparameterized restricted open-shell version, dubbed RO-B2PLYP, for the treatment of open-shell systems.[18] The very recently published DSD-BLYP functional by Kozuch et al. is a spin-component-scaled variant of the B2(GP)PLYP approaches, including the DFT-D dispersion correction.[20] Other recently developed DHDFs are the long-range corrected ωB97X-2 of Chai and Head-Gordon,[19] the XYG3 method of Zhang et al.,[21] and its modified versions XYG3s[22] and XYG3o.[23] An overview of the different double-hybrid approaches is given in Table 2.

In this study, we will investigate the B2PLYP, B2GPPLYP, DSD-BLYP, and XYG3 DHDFs together with our newly proposed methods PTPSS and PWPB95.

3.1. The B2PLYP and B2GPPLYP Functionals. B2PLYP and B2GPPLYP follow the same idea and just differ by amounts of Fock-exchange and perturbative correlation mixing. The first step in a B2(GP)-PLYP calculation is the generation of Kohn−Sham orbitals from the hybrid-GGA portion of the DHDF, which is denoted B2LYP or B2GPLYP (eq 2).

$$E_{\mathrm{XC}}^{\mathrm{B2(GP)LYP}} = (1 - a_{\mathrm{X}})E_{\mathrm{X}}^{\mathrm{B88}} + a_{\mathrm{X}}E_{\mathrm{X}}^{\mathrm{HF}} + (1 - a_{\mathrm{C}})E_{\mathrm{C}}^{\mathrm{LYP}} \tag{2}$$

The hybrid-GGA part contains the Becke 1988 (B88)[73] exchange $E_{\mathrm{X}}^{\mathrm{B88}}$ combined with nonlocal Fock-exchange $E_{\mathrm{X}}^{\mathrm{HF}}$ and Lee−Yang−Parr (LYP)[74,75] correlation $E_{\mathrm{C}}^{\mathrm{LYP}}$. The a_{X} and a_{C} are mixing parameters for the "exact" Fock-exchange and perturbative correlation, respectively. A second-order perturbation treatment (PT2), based on the KS orbitals and eigenvalues resulting from the B2(GP)LYP calculation, is carried out yielding the correlation energy $E_{\mathrm{C}}^{\mathrm{PT2}}$ that is scaled by the mixing parameter a_{C}. Although Brillouin's theorem is not valid, only double excitations from the KS determinant are considered. Single contributions are neglected and indirectly accounted for by the fitted scaling parameters. Thus, the final form of the B2(GP)PLYP exchange correlation energy is given by

$$E_{\mathrm{XC}}^{\mathrm{B2(GP)PLYP}} = E_{\mathrm{XC}}^{\mathrm{B2(GP)-LYP}} + a_{\mathrm{C}}E_{\mathrm{C}}^{\mathrm{PT2}} \tag{3}$$

In the case of B2PLYP, the two mixing parameters were fitted to the heats of formation (HOFs) of the G2/97 set by

Table 2. Overview of Various Double-Hybrid Density Functionals

functional	description	ref
B2PLYP	B88 exchange; LYP correlation; PT2 correlation based on hybrid-GGA part	ref 7
mPW2PLYP	like B2PLYP, but with mPW exchange	ref 13
B2KPLYP	reparameterized B2PLYP version for kinetics	ref 14
B2TPLYP	reparameterized B2PLYP version for thermochemistry	ref 14
B2GPPLYP	reparameterized B2PLYP version for general purpose applications	ref 15
B2πPLYP	reparameterized B2PLYP version for conjugated π-systems	ref 16
B2P3LYP	modified B2PLYP version with long-range PT2 correction	ref 17
B2OS3LYP	similar to B2P3LYP, but with SOS-PT2 correlation	
ROB2PLYP	reparameterized B2PLYP version within an ROKS formalism for treating open-shell systems	ref 18
ωB97X-2	ingredients of the B97 functional; long-range corrected; SCS-PT2 correlation	ref 19
XYG3	B88 exchange; LYP correlation; evaluated with B3LYP orbitals and densities	ref 21
XYG3s/XYG3o	modified XYG3 versions to account for basis set incompleteness	refs 22 and 23
DSD-BLYP	modified B2PLYP version with SCS-PT2 correction; fitted together with DFT-D dispersion correction	ref 20
PTPSS	reoptimized TPSS exchange and correlation; SOS-PT2 correlation; fitted together with DFT-D3 dispersion correction	this work
PWPB95	reoptimized PW exchange and B95 correlation; SOS-PT2 correlation; fitted together with DFT-D3 dispersion correction	this work

using a basis of quadruple-ζ quality (QZV3P). The parameters are $a_X = 0.53$ and $a_C = 0.27$. The parameters of B2GPPLYP were determined after taking into consideration atomization energies and reaction barrier heights by using the aug-pc2 and aug-pc3 basis sets. B2GPPLYP contains larger amounts of Fock-exchange ($a_X = 0.65$) and perturbative correlation ($a_C = 0.36$) than B2PLYP. It has been noted that these two double hybrids still lack an asymptotically correct description of long-range London-dispersion effects ($a_C < 1$), although the inclusion of the nonlocal PT2 part already leads to a qualitatively better description compared to common DFs. Therefore, it was suggested to combine the functionals with an empirical London-dispersion correction (DFT-D).[24]

Very recently, our group proposed a new version of this correction called DFT-D3, and we will make use of it in the present study.[43] Compared to the previously published versions,[72,76] DFT-D3 contains more "ab initio" ingredients and is characterized by less empiricism. It also contains system-specific C_6 and C_8 parameters, depending on the coordination sphere of each atom within a molecule. More details about this correction can be found in ref 43. In the present context, it is only necessary to mention that the dispersion correction E_{disp} includes two atom-pairwise terms:

$$E_{disp} = -\sum_{AB}\left(s_6 f_{d,6}(R_{AB}, s_{r,6})\frac{C_6^{AB}}{R_{AB}^6} + s_8 f_{d,8}(R_{AB})\frac{C_8^{AB}}{R_{AB}^8}\right) \quad (4)$$

where R_{AB} is the distance between two atoms A and B in a chemical system. The asymptotically relevant dipole−dipole term is scaled by a parameter s_6 and additionally contains a second parameter $s_{r,6}$, which scales cutoff radii within the damping function $f_{d,6}$. The second term is proportional to R_{AB}^{-8} and scaled by a factor s_8. For common DFs, s_6 is set to unity to ensure that the DFT-D3 correction has a physically correct asymptotic behavior. The other two parameters are fitted to a set of 130 noncovalent interaction energies. For double hybrids, an s_6 value smaller than unity has to be chosen, because of the presence of the nonlocal PT2 contribution. For B2PLYP, s_6 was originally set to 0.5.[77]

However, herein, we introduce a new scheme to estimate the s_6 value. We consider the three rare gas dimers Ne_2, Ar_2, and Kr_2 at large distances at which only long-range dispersion plays a role (7 Å for the neon dimer and 10 Å for the other two systems). The dispersion energies [i.e., $E_{disp}^{CCSD(T)} = E_{corr}^{CCSD(T)}(dimer) - 2E_{corr}^{CCSD(T)}(monomer)$] for these systems were then estimated at the CCSD(T)[78]/aug-cc-pVTZ[79] level of theory (carried out with Molpro 2009.1[80]), for which an asymptotically correct R_{AB}^{-6} behavior is expected. With the same basis set, the scaled perturbative dispersion energy of the considered DHDF (E_{disp}^{PT2}) is computed and compared to that of the coupled cluster treatment as the ratio between both dispersion energies ($E_{disp}^{PT2}/E_{disp}^{CCSD(T)}$). Finally, the average is taken over the three systems. To obtain the actual s_6 value, this average is subtracted from unity. Thus, an ideal method that correctly describes long-range dispersion interactions should have an s_6 of zero.

To validate our approach, we made some preliminary checks with the MP2, SCS-MP2, and SOS-MP2 methods. The MP2 method underestimates the CCSD(T) energy by 15% for Ne_2 but overestimated it by 10 and 18% for Ar_2 and Kr_2 (see Table S1 in the Supporting Information). On average, it gives a slight overestimation in the asymptotic limit, and thus the s_6 value is by −0.04 slightly negative. This overestimation is in accordance with previous observations (see, e.g., refs 20, 81, 82). The SCS-MP2 and SOS-MP2 methods underestimate the dispersion energies for all dimers and have s_6 values of 0.18 and 0.30, respectively. If the spin-opposite scale parameter of SOS-MP2 is set to unity, we obtain $s_6 = 0.47$, which is close to the expected value of one-half (in the asymptotic range, both the same and opposite spin parts have the same contributions).

After validating our approach, the s_6 value of B2PLYP was determined to be 0.64, which is larger than originally proposed. The other two parameters $s_{r,6}$ and s_8 were then refitted as described above and in the DFT-D3 paper. Following the same procedure, the s_6 value of B2GPPLYP turned out to be 0.56. The resulting values for $s_{r,6}$ and s_8 for both functionals are given in Table 3. More information about the results for the fit set used to determine the parameters can be found on our Web site.[51] It is recommended to use these revised parameters in future B2(GP)PLYP-D3 applications.

3.2. The DSD-BLYP Functional. The very recently published DSD-BLYP[20] functional is closely related to the

Table 3. Parameters for the DFT-D3 Correction

method	s_6	$s_{r,6}$	s_8
B2PLYP	0.64	1.427	1.022
B2GPPLYP	0.56	1.586	0.760
DSD-BLYP	0.50	1.569	0.705
PTPSS	0.75	1.541	0.879
PWPB95	0.82	1.557	0.705

B2(GP)PLYP approaches. It also contains B88 exchange and LYP correlation. However, the perturbative part is now based on the spin-component scaling idea (SCS-PT2).[46] The two scaling parameters (c_O and c_S) for the opposite ($E_C^{OS\text{-}PT2}$) and same-spin contributions ($E_C^{SS\text{-}PT2}$) are, moreover, independent from the scaling parameter (c_C) of the LYP correlation portion:

$$E_{XC}^{DSD-BLYP} = E_X^{B88} + a_X E_X^{HF} + c_C E_C^{LYP} + c_O E_C^{OS-PT2} + c_S E_C^{SS-PT2} \quad (5)$$

The three correlation scaling parameters and the amount of Fock-exchange were determined with a training set covering atomization energies, reaction barriers, noncovalently bound systems, transition metal compounds, and the MB08-186 set, which is also part of GMTKN30. Various basis sets of triple- and quadruple-ζ quality were used for this purpose. With an a_X value of 0.69, DSD-BLYP contains even more Fock-exchange than B2GPPLYP. The three correlation scale parameters are $c_C = 0.54$, $c_O = 0.46$, and $c_S = 0.37$. During the fitting procedure, the old DFT-D correction was applied, and the s_6 value was also fitted. In this work, we will make use of the DSD-BLYP functional, without changing the parameters. However, we will apply the new DFT-D3 correction and will refer to this combination as DSD-BLYP-D3. The three parameters were determined as described above for B2(GP)PLYP and are given in Table 3.

3.3. The XYG3 Functional. The XYG3 functional represents a different kind of B2PLYP variant. Instead of self-consistently creating the KS orbitals from the DHDF's hybrid-GGA part, the authors proposed to carry out first a normal B3LYP calculation. The resulting orbitals and density are used to evaluate both the empirically adjusted hybrid-GGA part (nonself-consistently) and the PT2 energy. When starting to work with the XYG3 functional, we had some problems with its implementation. We followed the author's description in the XYG3 paper.[21] According to them, the XYG3 formula reads

$$E_{XC}^{XYG3} = E_X^S + E_C^{VWN} + a_X(E_X^{HF} - E_X^S) + a_0\Delta E_X^{B88} + a_C(E_C^{PT2} - E_C^{LYP}) + (1 - a_C)\Delta E_C^{LYP} \quad (6)$$

where E_X^{Slater} stands for Slater exchange,[83] E_C^{VWN} for the VWN-LDA correlation,[84] ΔE_X^{B88} for the gradient correction part of the B88 functional, and ΔE_C^{LYP} for the gradient correction part of LYP [note that in later publications the factor $(1 - a_C)$ is missing before ΔE_{LYP}].[22,23,50] The reason for our problems seems to be the 100% of VWN correlation that is mixed in. In fact, we were not able to reproduce the results published in the XYG3 paper. However, we found another description of the XYG3 functional in a recent study by Vázquez-Mayagoitia et al.[85] that does not include VWN correlation. With that description, we were able to reproduce the results of the original XYG3 paper. This, apparently correct, formula reads

$$E_{XC}^{XYG3} = a_X E_X^{HF} + (1 - a_X)E_X^S + a_0\Delta E_X^{B88} + (1 - a_C)E_C^{LYP} + a_C E_C^{PT2} \quad (7)$$

In principle, XYG3 contains the same ingredients as B2PLYP and B2GPPLYP, with the exception that a different scaling of the LDA and semilocal exchange parts is applied. The three scale parameters a_X, a_0, and a_C were determined by a fit to the thermochemical data in the G3/99 set by applying the 6-311+G(3df,2p) basis ($a_X = 0.8033$, $a_0 = 0.2107$, and $a_C = 0.3211$). Thus, XYG3 is the DHDF with the largest amount of Fock-exchange (80.33%). We will later comment on the magnitude of the PT2 part in XYG3. Because the orbitals and density result from a different functional (e.g., B3LYP in XYG3) than the final semilocal exchange-correlation parts, XYG3 also contains an additional empirical degree of freedom compared to the other DHDFs.

3.4. The PTPSS Functional. The herein proposed PTPSS density functional ("P" stands for "perturbative") differs basically in four ways from B2PLYP and related methods. First of all, the key ingredients, i.e., the semilocal DFT parts, are changed from B88 exchange and LYP correlation to TPSS[86] exchange and correlation. Thus, PTPSS is a double-hybrid-meta-GGA functional given by

$$E_{XC}^{PTPSS} = (1 - a_X)E_X^{TPSS} + a_X E_X^{HF} + (1 - a_C)E_C^{TPSS} + a_C E_C^{OS-PT2} \quad (8)$$

The second major difference is that only contributions of electron pairs with opposite spin (OS) are included for the perturbative part $E_C^{OS\text{-}PT2}$, similar to the B2OS3LYP functional. This brings the formal scaling of N^5 with system size down to N^4, due to a Laplace transformation algorithm,[48] as first shown for the SOS-MP2 method.[47] We observed that by just neglecting the same spin (SS) terms, the errors for our fit set (see below) were reduced drastically. It is important to note here that PTPSS is only competitive at this SOS-PT2 level, while this is different with the B88/LYP parts where the same spin part must be included. This observation is consistent with the fact that (a) LYP does not contain any same-spin correlation, meaning that it must be considered by PT2 in a LYP-based DHDF, and (b) the same-spin correlation energy is not as accurate as the OS contribution at second order, as indicated by the success of SCS-MP2.[46] Thus, when the same spin part is already described well at the semilocal level by, e.g., TPSS, it seems better to neglect it in the PT2 treatment also to avoid double-counting effects entirely.

The third difference is that in previously published DHDFs only the scale factors for the Fock-exchange and perturbative correlation were fitted, whereas the semilocal DF parameters (e.g., β in B88) remained unchanged. Very recently, we discovered that the results with the TPSS ansatz improve significantly and become comparable to some results of

hybrid-GGA DFs when the seven parameters are refitted (termed oTPSS, where the prefix "o" stands for "optimized").[35] This observation inspired us in the first place to develop a double-hybrid based on TPSS, for which also these seven parameters are adjusted.

The TPSS exchange functional has the following form:

$$E_X^{TPSS} = \sum_\sigma \int \rho \varepsilon_X^{LDA} F_X^{TPSS} \, dr \tag{9}$$

in which ε_X^{LDA} is the LDA exchange energy density, σ is the spin variable (for α and β spin, respectively), and F_X^{TPSS} is the TPSS enhancement factor

$$F_X^{TPSS} = 1 + \kappa - \frac{\kappa}{1 + \frac{x}{\kappa}} \tag{10}$$

κ is the first functional parameter that is adjusted for PTPSS. The variable x is given as

$$\begin{aligned} x = \; & \left\{ \left[\frac{10}{81} + c\frac{z^2}{(1+z^2)^2} \right] p + \frac{146}{2025}\tilde{q}_b^2 - \right. \\ & \frac{73}{405}\tilde{q}_b \sqrt{\frac{1}{2}\left(\frac{3}{5}z\right)^2 + \frac{1}{2}p^2} + \frac{1}{\kappa}\left(\frac{10}{81}\right)^2 p^2 + \\ & \left. 2\sqrt{e}\frac{10}{81}\left(\frac{3}{5}z\right)^2 + e\mu p^3 \right\} / [(1 + \sqrt{e}p)^2] \end{aligned} \tag{11}$$

where

$$\begin{aligned} z &= \frac{\tau^W}{\tau} \quad \alpha = \frac{\tau - \tau^W}{\tau^{UEG}} \\ \tilde{q}_b &= \frac{9}{20}\frac{(\alpha - 1)}{[1 + b\alpha(\alpha - 1)]^{1/2}} + \frac{2p}{3} \\ p &= s^2 = \left(\frac{|\nabla \rho_\sigma|}{\rho_\sigma^{4/3} 2(3\pi^2)^{1/3}} \right)^2 \end{aligned} \tag{12}$$

ρ is the electron density, p is the square of the reduced spin variable s, τ is the kinetic energy density, τ^W is the von Weizsäcker kinetic energy density, and τ^{UEG} is the uniform gas kinetic energy density. μ, b, c, and e are four additional functional parameters that were adjusted for PTPSS.

The TPSS correlation functional is a modification of the correlation part of the Perdew−Kurth−Zapan−Blaha (PKZB) meta-GGA functional[87] and defined as follows

$$E_C^{TPSS} = \int \rho \varepsilon_C^{revPKZB} \times \left[1 + d\varepsilon_C^{revPKZB} \left(\frac{\tau^W}{\tau} \right)^3 \right] d^3r \tag{13}$$

with

$$\begin{aligned} \varepsilon_C^{revPKZB} = \; & \varepsilon_C^{PBE}[\rho_\alpha, \rho_\beta, \nabla\rho_\alpha, \nabla\rho_\beta][1 + C(\zeta, \xi)\left(\frac{\tau^W}{\tau}\right)^2] - \\ & [1 + C(\zeta, \xi)]\left(\frac{\tau^W}{\tau}\right)^2 \sum_\sigma \frac{\rho_\sigma}{\rho} \tilde{\varepsilon}_C \end{aligned} \tag{14}$$

and where

$$\begin{aligned} \tilde{\varepsilon}_C &= \max[\varepsilon_C^{PBE}[\rho_\sigma, 0, \nabla\rho_\sigma, 0], \varepsilon_C^{PBE}[\rho_\alpha, \rho_\beta, \nabla\rho_\alpha, \nabla\rho_\beta]] \\ C(\zeta, \xi) &= \frac{0.53 + 0.87\zeta^2 + 0.50\zeta^4 + 2.26\zeta^6}{\left[1 + \xi^2 \frac{(1+\zeta)^{-4/3} + (1-\zeta)^{-4/3}}{2} \right]^4} \\ \xi &= \frac{|\nabla\zeta|}{2(3\pi^2\rho)^{1/3}} \end{aligned} \tag{15}$$

Here, $\zeta = (\rho_\alpha - \rho_\beta)/\rho$ is the relative spin polarization. The TPSS correlation part depends on two parameters. These are d, as given in eq 13, and β, which is part of the PBE correlation functional[88] ε_C^{PBE} and the modified $\tilde{\varepsilon}_C$ (ref 86 gives a detailed description of all the necessary variables in the TPSS functional).

The values of the seven parameters of the original TPSS functional are given in Table 4 in comparison with the refitted values of oTPSS. We note in passing that, in 2007, Perdew et al. also published a reparameterized version of TPSS with different values for μ, c, and e[89] (for a redesigned version termed revTPSS, see ref 90).

The fourth major difference between PTPSS and most of the preceding DHDFs regards the fitting procedure. Like for oTPSS, B97-D,[72] or DSD-BLYP, PTPSS is fitted in combination with an empirical London-dispersion correction. Here, we applied the new DFT-D3 scheme.

The fitting procedure was carried out as follows: First of all, the amount of Fock-exchange was set to $a_X = 0.5$ and kept constant. We regard this value as a reasonable compromise for both main group and transition metal chemistry. We think that a too high fraction of Fock-exchange (e.g., 69% or about 80%, as in DSD-BLYP and XYG3, respectively) makes any DHDF unstable in electronically complicated situations. Evidence for this was already provided for the B2KPLYP functional in ref 15. The seven TPSS parameters and the SOS-PT2 parameter a_C were fitted in a standard least-squares procedure. The fit set (dubbed DFT fit set) is a modified version of the one we already used for oTPSS.[35] It is comprised of a total of 112 energies. These are 49 atomization energies (47 of the G2/97 set and additionally the adamantane and anthracene molecules, which are of a similar size but whose uniformly accurate description is difficult to achieve with DFs), five total atomic energies, eight atomic ionization potentials, and seven atomic electron affinities (taken from the G2-1 set), six noncovalently bound systems from the S22 set, the $(H_2O)_6$ cyclic cluster taken from WATER27, four rare gas dimers from RG6, and 29 decomposition energies from the MB08-165 benchmark set. Furthermore, the isomerization reaction from iso- to *n*-octane, the Diels−Alder reaction between furane and maleic anhydride to form the *endo* product, and the decomposition of Li_8 into lithium dimers (from ALK6) were included. The systems were weighted with different factors for the statistical analysis. The full set and the weight factors are listed in Table S2 in the Supporting Information.

During the fitting procedure, the DFT-D3 parameters were also adjusted. In a prescreening process, we used an s_6 value of 0.5. Later, we readjusted it, as described in section 3.1. The resulting value is 0.75, which is expected, as the functional contains less perturbative correlation than others,

Table 4. Parameters of the TPSS,[86] oTPSS,[34] and PTPSS Methods

	b	c	e	μ	κ	β	d	a_X	a_C
TPSS	0.40	1.59096	1.537	0.21952	0.804	0.06672	2.8		
oTPSS	3.43	0.75896	0.165	0.41567	0.778	0.08861	0.7		
PTPSS	0.15	0.88491	0.047	0.16952	0.872	0.06080	6.3	0.50	0.375

Table 5. Parameters of the PW,[91] mPW,[94] B95,[92] PW6B95,[93] and PWPB95 Methods

	b_{PW}	c_{PW}	d_{PW}	c_{opp}	$c_{\sigma\sigma}$	a_X	a_C
PW	0.0042	1.6455	4				
mPW	0.00426	1.6455	3.72				
B95				0.0031	0.038		
PW6B95	0.00538	1.7382	3.8901	0.00262	0.03668	0.28	
PWPB95	0.00444	0.3262	3.7868	0.00250	0.03241	0.50	0.269

due to the neglect of the same spin part. The $s_{r,6}$ and s_8 parameters were determined as described in the DFT-D3 paper (least-squares fit of a special van-der-Waals (vdW) fit set). Technically, we performed some fitting cycles for the electronic PTPSS parameters, then adjusted the DFT-D3 parameters and repeated this several times. The finally obtained parameter values are given in Tables 3 and 4. For these values, PTPSS-D3 yielded a root-mean-square deviation (RMSD) of 3.1 kcal/mol for the DFT fit set and 0.50 kcal/mol for the vdW fit set. More information about results for subsets of the vdW set are shown on our Web site.[51] Compared to TPSS and oTPSS, the parameters significantly change. Also, there are no obvious trends seen when comparing TPSS to oTPSS and TPSS to PTPSS. Particularly, the parameters b and e become very small, whereas d shows a large increase. The SOS-PT2 contribution in PTPSS is $a_C = 0.375$.

3.5. The PWPB95 Functional. A second new DHDF approach is dubbed PWPB95-D3. It is based on the Perdew−Wang (PW) GGA-exchange[91] and the Becke95 (B95) meta-GGA-correlation[92] functionals (inspired by Zhao and Truhlar's PW6B95 hybrid-meta-GGA[93]). PW exchange (E_X^{PW}) contains three adjustable parameters b_{PW}, c_{PW}, and d_{PW}.

$$E_X^{PW} = E_X^{LDA} - \sum_\sigma \int \rho_\sigma^{4/3} \frac{b_{PW} x_\sigma^2 - (b_{PW} - \beta) x_\sigma^2 \exp(-c_{PW} x_\sigma^2) - 10^{-6} x_\sigma^{d_{PW}}}{1 + 6 b_{PW} x_\sigma \sin h^{-1} x_\sigma - \frac{10^{-6} x_\sigma^{d_{PW}}}{A_x}} d^3 r \quad (16)$$

with

$$x_\sigma = \frac{|\nabla \rho_\sigma|}{\rho_\sigma^{4/3}}, \quad \beta = 5(36\pi)^{-5/3}, \quad A_x = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3} \quad (17)$$

The original parameter values for PW, the modified mPW,[94] and the reparameterized PW6B95 functional are shown in Table 5.

The B95 correlation functional can be divided into one part treating electron pairs of opposite spin (E_C^{opp}) and another one for those of same spin ($E_C^{\sigma\sigma}$):

$$E_C^{B95} = E_C^{B95,opp} + \sum_\sigma E_C^{B95,\sigma\sigma} \quad (18)$$

with

$$E_C^{B95,opp} = \frac{E_C^{PW}(\rho_\alpha, \rho_\beta) - E_C^{PW}(\rho_\alpha, 0) - E_C^{PW}(0, \rho_\beta)}{(1 + c_{opp} \sum_\sigma x_\sigma^2)} \quad (19)$$

$$E_C^{B95,\sigma\sigma} = \frac{2(\tau_\sigma - \tau_\sigma^W) E_C^{PW}(\rho_\sigma, 0)}{\frac{3}{5}(6\pi^2)^{2/3} \rho_\sigma^{5/3} (1 + c_{\sigma\sigma} x_\sigma^2)^2} \quad (20)$$

where E_C^{PW} is Perdew and Wang's correlation LSDA functional.[95] B95 correlation depends on two adjustable parameters c_{opp} and $c_{\sigma\sigma}$. The original B95 and the reparameterized PW6B95 values are also shown in Table 5.

Similar as in PTPSS-D3, the inherent functional parameters were refitted for PWPB95. PWPB95 also includes an SOS-PT2 correlation term. Due to the fact that B95 differs between a same and an opposite spin contribution, two different approaches for the PWPB95 functional are possible.

In the first one, the entire reparameterized B95 functional is scaled down by $1 - a_C$, where a_C is the scale parameter for the opposite spin perturbative contribution:

$$E_{XC}^{PWPB95} = (1 - a_X) E_X^{PW} + a_X E_X^{HF} + (1 - a_C) E_C^{B95} + a_C E_C^{OS-PT2} \quad (21)$$

This approach is in complete analogy with the PTPSS functional.

A second possibility is to include 100% of reparameterized, same-spin B95 correlation and to just scale down the opposite spin part by $1 - a_C$:

$$E_{XC}^{PWPB95} = (1 - a_X) E_X^{PW} + a_X E_X^{HF} + \sum_\sigma E_C^{B95,\sigma\sigma} + (1 - a_C) E_C^{B95,opp} + a_C E_C^{OS-PT2} \quad (22)$$

We considered both approaches and thoroughly compared them with each other. Our findings showed that the first approach yielded much better results for GMTKN30 than the second one. Thus, we will from now on refer to eq 21 whenever we discuss PWPB95.

PWPB95-D3 contains two adjustable parameters less than PTPSS-D3. Preliminary tests made it necessary to modify the fitting procedure compared to PTPSS-D3. Only one fit set was used with higher weights on noncovalently bound complexes. The DFT-D3 parameters ($s_{r,6}$, s_8) were fitted at the same time as the five functional parameters and the scale parameter a_C. Details about the fit set can be found in Table S3 of the Supporting Information. The amount of Fock-exchange was also fixed to 50%. The resulting RMSD value is 2.6 kcal/mol for the fit set compared to 3.6 kcal/mol for the alternative PWPB95-D3 approach according to eq 22. The finally obtained parameter values are given in Tables 3

and 5. All DFT parameters are smaller than for PW6B95. The parameter b_{PW} is significantly smaller than for all other functionals based on PW exchange (note that this was also observed for the reparameterized oPWLYP GGA-functional in ref 35). The SOS-PT2 contribution is, with $a_C = 0.269$, smaller than for PTPSS-D3. The s_6 value was adjusted as for the other DHDFs and is, as expected, larger than for PTPSS-D3 ($s_6 = 0.82$).

4. Computational Details

All calculations were carried out with a modified version of TURBOMOLE 5.9 and the original version of TURBOMOLE 6.0.[96−99] For the GMTKN30 analysis and the fitting procedures, the large Ahlrichs' type quadruple-ζ basis sets def2-QZVP were applied,[100] which yield results quite close to the Kohn−Sham limit. For the calculations of electron affinities, diffuse s and p functions (for hydrogen, only an s function) were added from the Dunning aug-cc-pVQZ basis sets;[79] the resulting set is denoted by aug-def2-QZVP. As discussed previously,[35] one diffuse s and one diffuse p function (taken from aug-cc-pVQZ) was added to oxygen in the case of WATER27. We also carried out calculations for GMTKN30 with the def2-TZVPP set and used diffuse functions from aug-cc-pVTZ where necessary. To account for scalar relativistic effects, the heavy atoms in HEAVY28 and RG6 were treated with the effective core potentials ECP-28 (for Sb−Xe) and ECP-60 (for Pb, Bi, and Rn),[101,102] which were slightly modified by Weigend and Ahlrichs and called "def2-ecp" in Turbomole.[100]

For all hybrid-(meta-)GGA parts of the DHDFs and for general hybrid functionals, the resolution of the identity (RI-JK) approximation was applied.[103] For the perturbative parts of the DHDFs, the RI approximation was used as well.[99] Auxiliary basis functions were taken from the TURBOMOLE basis set library.[104,105] In all cases, SCF convergence criteria were set to 10^{-7} E_h. For GMTKN30 calculations, the TURBOMOLE grid *m4* was used, whereas the larger *m5* grid was chosen in the fitting procedure.[105] All open-shell calculations were done within the unrestricted Kohn−Sham formalism (UKS). The study of GMTKN30 was carried out with the DHDFs B2PLYP,[7] B2GPPLYP,[15] DSD-BLYP,[20] XYG3,[21] PTPSS, and PWPB95 and as a comparison with the hybrids B3LYP[106,107] and PW6B95.[93] In all cases except for XYG3, the new empirical London dispersion correction (DFT-D3)[43] was applied, for which a separate program was used, which is available for download from our Web site.[51]

The study of dissociation reactions of transition metal carbonyls is based on a publication by Hyla-Kryspin and Grimme.[108] To make our evaluation consistent with this previous publication, we followed the same technical procedures. All calculations were carried out with a triple-ζ Gaussian basis augmented with polarization functions: (17s11p6d1f)/[6s4p3d1] for the transition metal and (11s6p2d)/[5s3p2d] for C and O.[109] Geometry optimizations were carried out with the BP86 functional.[73,110] Reaction enthalpies for the dissociation reactions (for more details, see section 5.25) were calculated as

$$\Delta H_R^{298} = \Delta E_{elec} + \Delta ZPVE + \Delta H_{0\rightarrow 298} + \Delta(PV) \quad (23)$$

where E_{elec} is the electronic energy, ZPVE is the zero point vibrational energy, $H_{0\rightarrow 298}$ is the thermal correction for a temperature of 298.15 K, and $\Delta(PV)$ is the (ideal gas) volume work. ΔZPVE and $\Delta H_{0\rightarrow 298}$ were calculated with the BP86 functional by using the SNF[111] program and applied for all tested functionals.

5. Results and Discussion

5.1. The GMTKN30 Database. *5.1.1. Results for (aug-)def2-QZVP.* The MADs and RMSDs for all functionals obtained with (aug-)def2-QZVP are shown in Tables S4−S11 in the Supporting Information. As both values provide the same conclusions, only MADs will be discussed in the following. First of all, the DHDFs are compared with functionals on the hybrid level. In our study, these are the B3LYP-D3 and PW6B95-D3 methods. We note in passing that the latter clearly outperforms B3LYP-D3 and is, according to our experience, overall one of the best hybrid functionals on the market.

The graphs in Figure 2 show the ratios between the MADs of the double-hybrids and the hybrids. This comparison confirms that all DHDFs outperform the hybrids. The only exceptions are the noncovalent interactions, for which B3LYP-D3 is significantly better than XYG3 for S22, ADIM6, RG6 (factor of 3), HEAVY28, PCONF, and ACONF. PTPSS-D3 is slightly worse than B3LYP-D3 for RG6, PCONF, and ACONF. B3LYP-D3 is better than PWPB95-D3 for HEAVY28. However, in these last four cases, the PTPSS-D3, PWPB95-D3, and B3LYP-D3 methods are already within the errors of the reference values, and thus, this comparison should be interpreted with care. The comparison between the DHDFs and PW6B95-D3 shows again that PW6B95-D3 is outperformed in most of the cases but that the ratios are closer to unity and in a narrower range than for B3LYP-D3. PW6B95-D3 is, like B3LYP-D3, better than XYG3 for noncovalent interactions. The high ratios for the RG6 set, though, are a bit misleading, as all DHDFs, except XYG3, yield excellent MADs for that set: 0.06 (B2PLYP-D3), 0.05 (B2GPPLYP-D3), 0.07 (DSD-BLYP-D3), 0.09 (PTPSS-D3), 0.05 (PWPB95-D3), and 0.20 kcal/mol (XYG3) compared to 0.03 kcal/mol (PW6B95-D3).

After having seen that mixing in a perturbative correlation generally improves the results compared to fourth-rung functionals, we concentrate now on a comparison of the results for the DHDFs. All MADs based on (aug-)-def2-QZVP calculations can be found in the left column of Figure 3. The MB08-165 set contains very difficult, randomly created molecular systems and is a very good indicator for the robustness of a quantum chemical method. B2PLYP-D3, B2GPPLYP-D3, and PTPSS-D3 are in a same range, with MADs of about 4 kcal/mol. XYG3 is significantly worse, with 5.2 kcal/mol (even the PW6B95-D3 hybrid is better with 4.7 kcal/mol). DSD-BLYP-D3 is a clear improvement, with 3.4 kcal/mol. PWPB95-D3 yields the best MAD ever reported for this test set (2.5 kcal/mol). This value is in the range of CCSD(T)/cc-pVQZ, that was reported to yield an MAD of 2.6 kcal/mol.[32] This finding can be

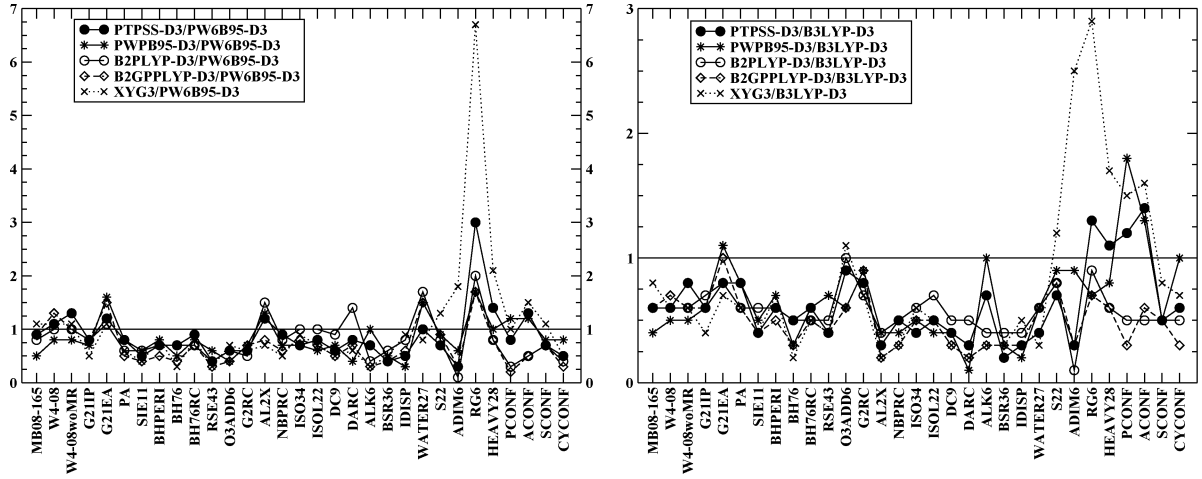


Figure 2. Ratios of the MADs of different functionals: comparisons between the DHDFs and PW6B95-D3 (left); comparisons between the DHDFs and B3LYP-D3 (right). All calculations were carried out with (aug)-def2-QZVP. To make the curves better distinguishable from each other, the curves for the DSD-BLYP functional were left out. The results are in qualitative agreement with those shown in the figures, though.

interpreted as a first hint on the robustness of this functional. PWPB95-D3 also gives the best MAD for the atomization energy test set (1.9 kcal/mol). The other five DHDFs are within a range of 2.4−3.0 kcal/mol and quite similar to each other. This similarity is in contrast to frequent claims in favor of XYG3 over B2PLYP, which are based on the calculations of HOFs.[41] Because theoretical HOFs are basically atomization energies, this discrepancy is at present not understandable for us. For ionization potentials, XYG3 yields the best MAD (1.4 kcal/mol), whereas the other DHDFs are in the range of about 2.0−2.3 kcal/mol. Electron affinities are almost equally well described by XYG3, B2PLYP-D3, and PTPSS-D3, whereas the description is slightly worse for B2GPPLYP-D3, DSD-BLYP-D3, and PWPB95-D3. A similar trend can also be seen for proton affinities, with the exception that PTPSS-D3 and PWPB95-D3 are slightly worse than the other four methods (note that here delocalized systems also play a role and that the delocalization error is expected to be larger the less Fock-exchange is included).

Larger differences in the MADs are observed for SIE-related problems. The SIE11 set is best described by XYG3 and DSD-BLYP-D3, with 3.1 kcal/mol, followed closely by B2GPPLYP-D3, with 3.4 kcal/mol. B2PLYP-D3 shows the worst MAD of the DHDFs, with 4.9 kcal/mol. This trend follows qualitatively the amount of Fock-exchange that decreases when going from XYG3 to DSD-BLYP, to B2GPPLYP, and to B2PLYP. However, PTPSS-D3 and PWPB95-D3 have the lowest amount of Fock-exchange of all DHDFs, and their MADs are, with 3.9 and 4.3 kcal/mol, better than for B2PLYP-D3. A possible explanation might be the single-electron SIE correction within the TPSS and B95 parts. The result for the barrier heights of the substitution, association, and unimolecular and transfer reactions within BH76 shows a slightly different picture. Here, XYG3, DSD-BLYP-D3, and B2GPPLYP-D3 are the best functionals with MADs of 1.1, 1.2, and 1.3 kcal/mol. This time, though, B2PLYP-D3 and PTPSS-D3 have the same MADs, with 2.5 kcal/mol. PWPB95-D3 gives 1.8 kcal/mol. However, this picture cannot be generalized to other barriers, particularly when larger (closed-shell) systems are involved. This is seen from the results for BHPERI, where XYG3 and PWPB95-D3 are the worst functionals, with 1.9 kcal/mol, followed by PTPSS-D3, with 1.7 kcal/mol, by B2PLYP-D3 (1.6 kcal/mol), by B2GPPLYP-D3 (1.3 kcal/mol), and by DSD-BLYP-D3 (1.2 kcal/mol).

The evaluation of reaction energies shows a heterogeneous picture, however, at a generally high level of accuracy. In some cases, the MADs of all six functionals are close to each other (e.g., for BH76RC), and in other cases, they differ significantly (for O3ADD6, AL2X, ISOL22, DC9, DARC, ALK6, and BSR36). However, no single DHDF is consistently the best one. XYG3 is the best functional in the case of AL2X and together with DSD-BLYP-D3 for NBPRC. DSD-BLYP-D3 is, moreover, the best method for DC9 and BSR36. B2GPPLYP-D3, PTPSS-D3, and PWPB95-D3 results are often very close, though. Sometimes, XYG3 can be also the worst functional, and it is comparable to B2PLYP-D3. The new PTPSS-D3 performs best for ISO34 (MAD = 0.9 kcal/mol). PWPB95-D3 is the best functional for O3ADD6 (1.7 kcal/mol, together with B2GPPLYP-D3), ISOL22 (2.9 kcal/mol), and DARC (1.5 kcal/mol). However, it has an outlier for ALK6 (4.6 kcal/mol), which is almost the same value as for B3LYP-D3 and PW6B95-D3 (4.7 kcal/mol).

The results for the noncovalent interactions test sets are more uniform. Intramolecular dispersion interactions within IDISP are best described by PWPB95-D3 (MAD = 1.2 kcal/mol), followed by DSD-BLYP-D3 (MAD = 1.4 kcal/mol), PTPSS-D3 (MAD = 1.7 kcal/mol), and B2GPPLYP-D3 (2.1 kcal/mol). B2PLYP-D3 and XYG3 have larger MADs with 3.0 and 3.1 kcal/mol. The water clusters in WATER27 are best described by XYG3, DSD-BLYP-D3, and PTPSS-D3 with 1.4, 1.5, and 1.6 kcal/mol. PWPB95-D3, B2GPPLYP-D,3 and B2PLYP-D3 have MADs of 2.4, 2.6, and 2.8 kcal/mol. The S22 set is best described by PTPSS-D3 with a very low MAD of 0.25 kcal/mol. It is followed by B2PLYP-D3 (0.27 kcal/mol), DSD-BLYP-D3 (0.28 kcal/mol), B2GPPLYP-D3 (0.30 kcal/mol), PWPB95-D3 (0.32 kcal/mol), and XYG3 (0.45 kcal/mol). Note that we did not use any London-dispersion correction for XYG3, as in recent publications it

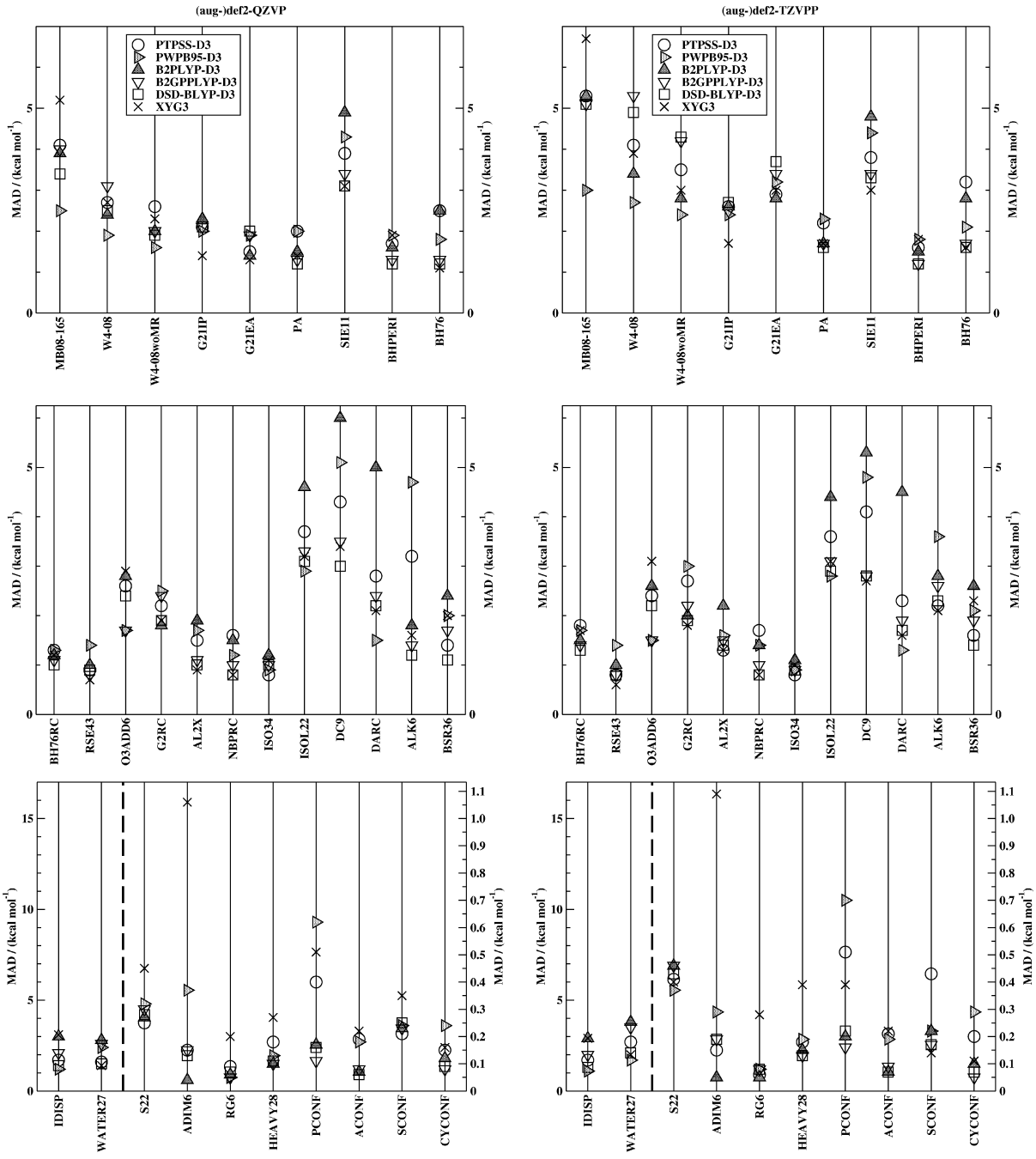


Figure 3. MADs for PTPSS-D3, PWPB95-D3, B2PLYP-D3, B2GPPLYP-D3, DSD-BLYP-D3, and XYG3 in kcal/mol for the complete GMTKN30 database with (aug-)def2-QZVP (left column) and (aug-)def2-TZVPP (right column). For the sets to the right of the dashed lines, the right MAD axes apply.

was argued that XYG3 works also very well for dispersion dominated interactions on its own.[21−23,50,85] Indeed, the overall behavior is not bad at first glance, but the comparison with the results for the other functionals shows that dispersion interactions are not fully considered in XYG3. A very prominent example is the ADIM6 test set, for which XYG3 yields a large MAD of 1.1 kcal/mol, whereas the other MADs are <0.15 kcal/mol, which is within the accuracy of the reference method (except for PWPB95-D3, with 0.36 kcal/mol). Also, for the other subsets (except for PCONF, where PTPSS-D3 and PWPB95-D3 have MADs of 0.40 and 0.62 kcal/mol), it can be seen that all DHDFs, except XYG3, are close to the accuracy of the estimated CCSD(T)/CBS reference values. Therefore, a direct comparison between these five functionals and a ranking of them is not appropriate. Dispersion corrections for XYG3 in its present form make no sense because double-counting effects cannot be avoided for such a highly nonlocal functional that has been parametrized without such corrections.

To better rationalize the results discussed above, one can simply count how many times a certain DHDF yields the best MAD or RMSD for a subset. The results are depicted in Figure 4. B2PLYP-D3 yields the best MAD in three, B2GPPLYP-D3 in five, DSD-BLYP-D3 in nine, XYG3 in eight, PTPSS-D3 in three, and PWPB95-D3 in seven cases. This picture allows a conclusion that strongly favors the

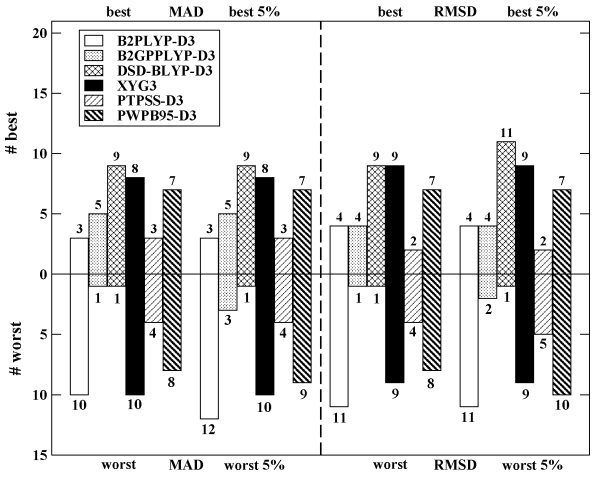


Figure 4. Analysis of how many times a certain double-hybrid yields the best and the worst MAD (left) or RMSD (right) for all subsets of GMTKN30 and how many times a functional is within a 5% range of the best/worst MAD or RMSD.

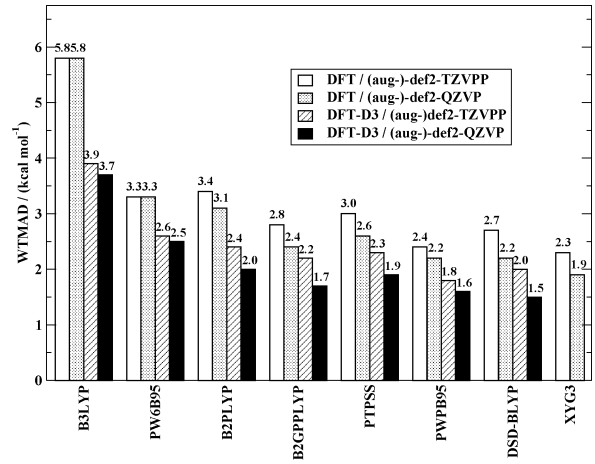


Figure 5. Weighted total mean absolute deviations (WTMADs) in kcal/mol for all tested methods with and without dispersion correction. The values are based on (aug-)def2-TZVPP and (aug-)def2-QZVP calculations.

DSD-BLYP-D3 and XYG3 functionals. However, if one also counts how many times a certain DHDF is the worst compared to the other DHDFs, one obtains a different picture. B2PLYP-D3 and XYG3 yield the worst MADs in 10 cases each. PWPB95-D3 follows with eight cases. PTPSS-D3 is the worst functional in only four cases, B2GPPLYP-D3 and DSD-BLYP-D3 in one case each. Thus, on the one hand, XYG3 seems to compete with the DSD-BLYP-D3 method, but on the other hand, it is also by far outperformed by DSD-BLYP-D3, B2GPPLYP-D3, and PTPSS-D3 and as bad as B2PLYP-D3. PWPB95-D3 shows a similar trend but is slightly better than XYG3 in this analysis. We think that, besides always obtaining the best MAD for a certain subset, it is also important that a functional is robust and shows a uniform accuracy for a whole range of different properties. Testing for robustness is the main reason why we created the GMTKN24/30 databases. Particularly, when being challenged by new chemical problems, it is in our opinion wiser to use a functional that is more robust. The results shown in Figure 4 give a hint on this property. The RMSDs on the right-hand side of Figure 4 allow the same conclusions. Besides just counting how many times the best or the worst MAD/RMSD is obtained, we also analyzed how many times a functional is within a 5% range of the best and worst result. These trends are comparable to the ones discussed above.

A third statistical evaluation of the results is carried out by considering the WTMADs (Figure 5). All six double hybrids have lower WTMADs than PW6B95-D3, the best hybrid. Again, we observe that the London-dispersion correction improves the results. B2PLYP-D3 has a WTMAD of 2.0 kcal/mol, which is the worst, compared to the other DHDFs. XYG3 is, by 0.1 kcal/mol, better (WTMAD = 1.9 kcal/mol). PTPSS-D3 yields the same WTMAD. B2GPPLYP-D3 is, with 1.7 kcal/mol, the third best; PWPB95-D3, with 1.6 kcal/mol, the second best; and DSD-BLYP-D3, with 1.5 kcal/mol, the best DHDF. This again indicates the better overall performance of DSD-BLYP-D3 and PWPB95-D3. The new PWPB95-D3 functional is thus a significant improvement over B2PLYP, XYG3 and B2GPPLYP and a good alternative to DSD-BLYP-D3, at a much lower computational cost for large systems and employing less nonlocality.

5.1.2. Basis Set Dependance. In section 3, it was discussed that the different DHDFs were developed by applying basis sets of different quality. The basis set dependances of the Fock-exchange and perturbative correlation parameters were already studied.[14,15,18] Furthermore, also, a justified question was raised as to whether a functional fitted with a triple-ζ basis set works better in practice with a basis set of similar quality than a functional developed with a quadruple-ζ basis.[21−23,50] This is particularly important, as in common applications, basis sets of quadruple-ζ quality, as used here, are not always feasible.

The GMTKN30 database allows us to answer this question on very solid ground. The right column of Figure 3 shows the MADs of all 30 subsets obtained with the (aug-)-def2-TZVPP basis (see also Tables S12−S19, Supporting Information). Figure 5 shows the WTMADs with and without dispersion correction for all tested methods. It can be seen that the basis set effect for the hybrid functionals is very small. The WTMADs increase by merely 0.2 and 0.1 kcal/mol for B3LYP-D3 and PW6B95-D3 (to 3.9 and 2.6 kcal/mol), and we can conclude that for these functionals the Kohn−Sham limit is almost reached with the (aug-)def2-TZVPP basis. The basis set dependence for most of the DHDFs is stronger, though. Usually, the MADs for the subsets worsen; sometimes they become slightly better. The stronger effect can be explained by the presence of the WF-based perturbative correction. Interestingly, DSD-BLYP-D3 and B2GPPLYP-D3 results lie closer to each other than on the quadruple-ζ level. This can, for example, be seen for MB08-165, where both functionals yield the same MADs of 5.1 kcal/mol.

The WTMADs of B2PLYP-D3 and XYG3 worsen by 0.4 kcal/mol compared to the quadruple-ζ level. B2GPPLYP-D3 and DSD-BLYP-D3 are even more affected (increase of 0.5 kcal/mol), which is in line with their large PT2 contributions. PTPSS-D3 gets worse by 0.4 kcal/mol. PWPB95-D3 shows the least basis set dependence (increase from 1.6 to

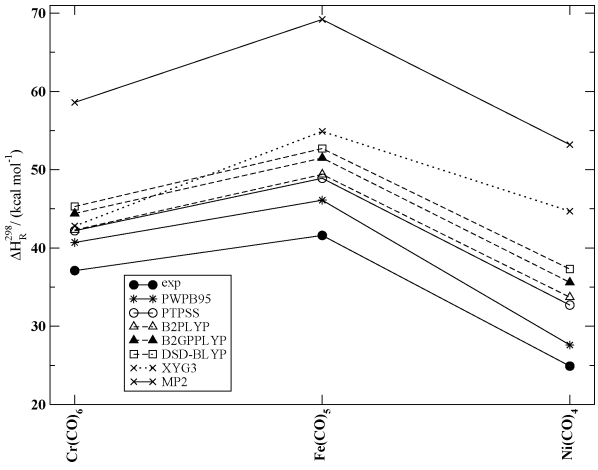


Figure 6. Reaction enthalpies (ΔH_R^{298}) in kcal/mol for the first CO dissociation reaction of three transition metal carbonyls. The values were obtained with the PWPB95, PTPSS, B2PLYP, B2GPPLYP, DSD-BLYP, and XYG3 methods. Vibrational and thermal corrections are based on BP86 calculations. A triple-ζ basis was applied in all cases [(17s11p6d1f)/[6s4p3d1] for the transition metal and (11s6p2d)/[5s3p2d] for C and O]. The experimental and MP2 values are taken from ref 108.

1.8 kcal/mol) and is, thus, the only DF that has a WTMAD below 2 kcal/mol at this basis set level. Thus, at the triple-ζ level, the order of accuracy changes. B2PLYP-D3 still has the largest WTMAD compared to the other DHDFs (2.4 kcal/mol). XYG3 and PTPSS-D3 have WTMADs of 2.3 kcal/mol and are followed closely by B2GPPLYP-D3 with 2.2 kcal/mol. DSD-BLYP-D3 is the second best method, with 2.0 kcal/mol. PWPB95-D3 is the best method, with 1.8 kcal/mol, although it was fitted with a quadruple-ζ basis. In fact, this also indicates a higher robustness of PWPB95-D3 compared to the other DHDFs. According to these results, the answer to the question, whether a functional fitted with a triple-ζ basis set gives relatively better results when applied with such a basis set, is clearly no and thus contrary to claims in the literature.[21−23,50] The WTMADs are just shifted, and the differences between XYG3 and B2PLYP-D3 are still the same compared to the quadruple-ζ results. Our findings, also contradict recent claims that the basis set dependence of XYG3 is similar to that of B3LYP.[23]

5.2. Performance for Transition Metal Carbonyls. In the theoretical section about the DHDFs, we argued, that we have chosen the Fock-exchange parameter to be one-half, so that both, main group and transition metal chemistry, can be described adequately. As a test of this hypothesis, we studied the dissociation of one CO ligand in the $Cr(CO)_6$, $Fe(CO)_5$, and $Ni(CO)_4$ isoelectronic series, which is sensitive to details of the correlation treatment. Reaction enthalpies were calculated as described in the computational details section and are shown in Figure 6. Experimental and MP2 values,[108] which were obtained with the same basis set and the same vibrational and thermal corrections as for the DHDFs, are also given (see also Table S20, Supporting Information).

All theoretical methods overestimate the experimental values. PWPB95, though, is closest to the reference, followed by PTPSS, B2PLYP, B2GPPLYP, DSD-BLYP, XYG3, and MP2. As anticipated, the amount of Fock-exchange and the size of the error are strongly related. Interestingly, the values for all methods, except for XYG3, are just shifted with respect to the experiment (the connecting lines in Figure 6 are almost parallel to the experimental reference). This, however, is not observed for XYG3, for which the chromium compound is relatively better described than the other two compounds and the dissociation energy of $Ni(CO)_4$ is incorrectly computed to be higher than for $Cr(CO)_6$.

5.3. A Comment on the XYG3 Functional. We have discussed many results for main group chemistry and given some insight into the functionals' performance for an exemplary transition metal reaction. We concluded that PWPB95 and PTPSS seem to be rather robust functionals and that also XYG3 seems to perform reasonably well, yielding good MADs in many cases. In this section, we further want to comment on XYG3 and give a new explanation for its good performance.

In their original XYG3 paper, the authors argue that the hybrid-GGA part of B2PLYP (i.e., B2LYP) does not employ 100% of DFT correlation and that this "truncated DFT" method yields densities and orbitals "that are dramatically different from the real ones." They argue that using B3LYP densities and orbitals for the evaluation of XYG3 is the key ingredient for the functional's good performance, because B3LYP densities had been shown to be "similar to those from CCSD(T) wave functions" and because B3LYP has 100% DFT correlation.

We do not agree with this interpretation. Of course, e.g., for thermochemical properties, some correlation is missing in B2LYP. This can be seen, when, e.g., the WTMADs for the GMTKN30 set for BHLYP[112] and B2LYP are compared with each other (6.7 vs 9.4 kcal/mol). However, the missing 27% LYP correlation does not affect significantly either the shape of the orbitals or the electron density. On the contrary, the results for electronic excited states (vertical excitation energies, oscillator and rotational strengths) are practically the same for TD-B2LYP and TD-BHLYP, as discussed by us several times.[36−38] The reason for this is the comparable one-particle spectrum for both methods. BHLYP has the same ingredients as B2LYP, including similar amounts of Fock-exchange, and in principle the two methods just differ by the amounts of LYP correlation. In our context, it is the amount of Fock-exchange that is the decisive factor for the orbital energies. In fact, the occupied part for normal molecules is rather similar for different functionals (and even overlaps to HF are large).[113] The reason for this is the exchange and not the correlation part! In fact, large differences are found for the virtual spectrum, which is worse with B3LYP due to the wrong asymptotic behavior of the larger (compared to B2LYP) GGA part. This can be also seen from theoretical electronic spectra, particularly including excitations into (diffuse) Rydberg orbitals.[114] Thus, we do not think that excitations to B3LYP virtual orbitals lead to better results in a perturbation theory because B3LYP has 100% DFT correlation. Instead, we tentatively assign part of the relatively high accuracy of XYG3 for main group problems to the small amount of Fock-exchange in the orbital generation step, which improves, e.g., spin-contaminated

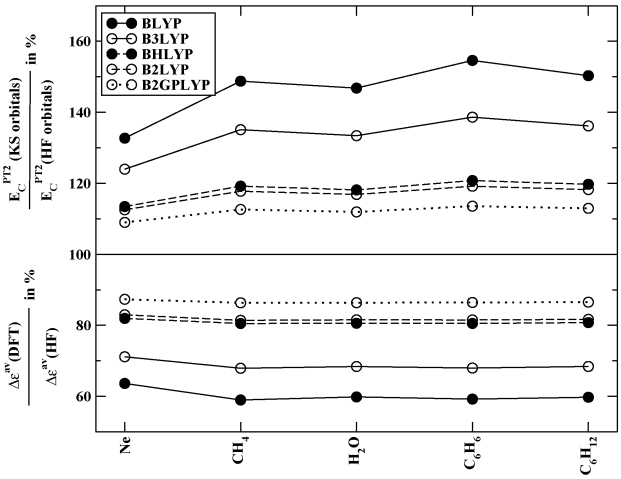


Figure 7. Percentage ratios of the averaged orbital energy gap ($\Delta\varepsilon^{av}$) of five molecules for different KS-DFT methods and HF theory (lower part) and ratios between unscaled PT2 correlation energies based on KS and HF orbitals (upper part). All calculations were carried out with def2-QZVP.

problems (see also ref 115 for a recent investigation that also compares orbital energies with the amount of Fock-exchange).

In the following, we would like to underline the similarities between BHLYP and B2LYP and would like to support the above statements by solid data. Therefore, we carried out SCF calculations for five examples (Ne, CH_4, H_2O, C_6H_6, and C_6H_{12}) with BLYP, B3LYP, BHLYP, B2LYP, B2GPLYP, and HF (def2-QZVP basis). We then calculated an average occupied−virtual orbital gap $\Delta\varepsilon^{av}$ for each method. We created for each system and each method all possible single excitations from the occupied valence orbitals into the corresponding valence virtual orbitals and then took the average over all excitation energies.

The actual energy values can be found in Table S21 (Supporting Information). We then related the $\Delta\varepsilon^{av}$(DFT) values obtained for the Kohn−Sham methods to $\Delta\varepsilon^{av}$(HF) obtained at the HF level. This is shown in the lower half of Figure 7. All average DF gaps are smaller than for HF. For each method, this underestimation seems to be rather system-independent. An inverse relation between the amount of Fock-exchange and $\Delta\varepsilon^{av}$ can be seen. BLYP yields only about 60% of the HF gap, and B2GPLYP gives the highest ratio of 87%. The results for B2LYP and BHLYP are very similar, proving the above statement, that only the amount of Fock-exchange is important for the one-particle spectrum and not the portion of LYP correlation. B3LYP has a lower average gap than B2LYP (68% vs 81%).

Furthermore, we carried out a standard second-order perturbative calculation based on the different KS orbitals and eigenvalues and related the resulting unscaled correlation energies to the canonical MP2 result (top part of Figure 7). As expected, the absolute correlation energies based on KS orbitals are higher than for MP2. The values obtained with BHLYP and B2LYP are very similar. The PT2 correlation energy is found to be inversely proportional to the amount of Fock-exchange as expected from the orbital energy differences entering the denominator of the PT2 energy expression.

As B3LYP yields a lower gap, a PT2 calculation with B3LYP orbitals yields a higher absolute correlation energy than for B2LYP orbitals. Consequently, the XYG3 functional includes a higher amount of "effective" PT2 correlation than B2LYP and B2GPLYP. This can be quantified by scaling the actual parameters a_C (which are 0.27, 0.36, and 0.3211 for the three functionals) by the average factor by which the MP2 correlation is overestimated, i.e., 1.169 for B2PLYP, 1.120 for B2GPPLYP, and 1.335 for XYG3 (Figure 7). These "effective" PT2 contributions are 32% for B2PLYP, 40% for B2GPPLYP, and 43% for XYG3. Thus, part of the explanation for the good behavior of XYG3 is not the fact that B3LYP orbitals are "better" than B2LYP ones due to 100% DFT correlation but that a PT2 treatment with B3LYP orbitals introduces an effectively higher amount of nonlocal correlation.

Furthermore, compared to the B2PLYP and B2GPPLYP, XYG3 has a very reduced amount of repulsive gradient corrected exchange contribution (ΔE_X^{B88}). This, in combination with the higher effective nonlocality, explains why in some studies on noncovalent interactions XYG3 yielded reasonable results.[21,22,50,85] Note, however, that this only holds for small- and medium-sized complexes as XYG3 misses asymptotically about 57% of the dispersion energy due to an effective a_C of 0.43.

Although decoupling the orbital/density generation and the actual functional evaluation seems to give reasonable results for ground state related problems, we are not sure whether this argument also holds for other properties. A time-dependent treatment of excited states within a TD-XYG3 formalism, for example, would lead to a not well-defined RPA-type matrix because of the different treatments of orbital energies and the functional's kernel. This is not the case for TD-B2PLYP and TD-B2GPPLYP.[36,38]

6. Conclusions

We presented an extension and modification of the recently published GMTKN24 database[35] for applications to general main group thermochemistry, kinetics, and noncovalent interactions. This extended and improved version is called GMTKN30 and comprises 30 different benchmark sets. In total, 1218 single point calculations have to be carried out to evaluate 841 data points (relative energies). In this study, we particularly focused on the analysis of double-hybrid density functionals (DHDFs) and presented two new DHDFs called PTPSS and PWPB95. PTPSS consists of semilocal TPSS exchange and correlation parts, for which seven inherent functional parameters were refitted. PWPB95 contains reparameterized PW exchange and B95 parts (five parameters). Both functionals mix in 50% of nonlocal Fock-exchange, which is less than for other DHDFs. PTPSS includes 37.5% and PWPB95 includes 26.9% of spin-opposite scaled second-order perturbative correlation (SOS-PT2). When combined with a Laplace transformation type algorithm, they scale only as $O(N^4)$ with system size. Furthermore, both methods are combined with the latest version of the empirical London-dispersion correction (DFT-D3),[43] for which we also presented a new scheme with which to estimate a DHDF's s_6 scaling parameter. PTPSS-D3 and

PWPB95-D3 were studied for the complete GMTKN30 database and dissociation reactions of three prototypical transition metal carbonyls. The results were compared with the hybrids B3LYP and PW6B95 and with the double hybrids B2PLYP, B2GPPLYP, DSD-BLYP, and XYG3. The analyses led to the following conclusions:

(1) All double hybrids clearly outperform the hybrid functionals. B2PLYP-D3, B2GPPLYP-D3, and PTPSS-D3 yield very good MADs (about 4 kcal/mol) for the difficult MB08-165 subset, for which XYG3 is even worse than the hybrid PW6B95-D3. DSD-BLYP-D3 is better with 3.4 kcal/mol. PWPB95-D3 yields an excellent result of 2.5 kcal/mol, which is better than CCSD(T)/cc-pVQZ results (2.6 kcal/mol). XYG3 is also worse for noncovalent interactions than the other dispersion-corrected DHDFs, and the errors can reach up to 1 kcal/mol (MAD for alkane dimers). The other five methods usually have errors, which are already within the accuracy of the respective reference data (MADs of less than about 0.2 kcal/mol). Over the whole GMTKN30 database, DSD-BLYP-D3 yields in nine cases the best MAD compared with the other double hybrids (eight for XYG3, seven for PWPB95-D3, five for B2GPPLYP-D3, and three for B2PLYP-D3 and PTPSS-D3). On the other hand, XYG3 and B2PLYP-D3 yield in 10 cases the worst MAD. PWPB95-D3 is slightly better (eight cases). PTPSS-D3 gives in four cases the worst MADs; B2GPPLYP-D3 and DSD-BLYP-D3 in one case each. For the application to new, hitherto unexplored chemical problems, the DSD-BLYP-D3 and PWPB95-D3 functionals seem to be a good choice. The WTMADs [for an (aug-)def2-QZVP basis] underline the general applicability of PWPB95-D3 and DSD-BLYP-D3: 1.6 kcal/mol for PWPB95-D3 and 1.5 kcal/mol for DSD-BLYP-D3. B2GPPLYP-D3 has a WTMAD of 1.7 kcal/mol. PTPSS-D3 and XYG3 have a WTMAD of 1.9 kcal/mol. B2PLYP-D3 follows with 2.0 kcal/mol.

(2) Further studies of GMTKN30 with the (aug-)def2-TZVPP basis reveal that, except for PWPB95-D3, DHDFs are (due to the perturbative correction) more basis set dependent than hybrids. The WTMADs are still better than for hybrids but also closer to them than for (aug-)def2-QZVP. Although fitted with a quadruple-ζ basis set, PWPB95-D3 is the best DHDF at the triple-ζ level, the least basis set dependent one (the dependence is similar to that of hybrids) and the only functional yielding a WTMAD below 2 kcal/mol. The WTMAD for PWPB95-D3 is 1.8 kcal/mol. It is followed by DSD-BLYP-D3 with 2.0 kcal/mol and by B2GPPLYP-D3 with 2.2 kcal/mol. PTPSS-D3 and XYG3 yield 2.3 kcal/mol. B2PLYP has a WTMAD of 2.4 kcal/mol.

(3) Calculated reaction enthalpies for the first CO dissociation in transition metal carbonyls by DHDFs and MP2 are larger than the experimental reference values. However, PWPB95-D3 is closest to experimental values with an error of about 3 kcal/mol (about 5−10% of D_e). This is attributed to the smallest amount of Fock-exchange compared to the other DHDFs. Only XYG3 fails to reproduce the trend of the dissociation enthalpies in the series of compounds.

(4) The good performance of XYG3 for main group compounds can be explained by a large effective local correlation contribution. In contrast to claims in the literature,[21−23,50] we find that the "better orbitals" that are used for the evaluation of XYG3 are more influenced by the smaller Fock-exchange in B3LYP and not by a full semilocal correlation part as claimed in refs 21−23 and 50. A study of different functionals and comparisons with HF proved that functionals with less Fock-exchange have lower average single-particle gaps. This leads to higher correlation energies if the resulting orbitals are used in the perturbative treatment. Thus, XYG3 has effectively a higher perturbative contribution than, e.g., B2PLYP and B2GPPLYP. This higher nonlocality and a reduced amount of the over-repulsive Becke 1988 gradient correction explains the good behavior of XYG3 for noncovalent interactions in previous studies, in which only medium-sized systems were considered.[21−23,50,85] Asymptotically, it misses about 57% of the dispersion energy and is thus not recommended in its present form for the calculation of van der Waals interactions in large systems.

In summary, our investigations revealed that PTPSS-D3 and PWPB95-D3 are valuable new functionals. They have a better formal scaling with system size than other DHDFs. They are robust for main group chemistry and seem to perform better than other DHDFs for transition metal compounds. PWPB95-D3 is the best DHDF for applications at a triple-ζ level. It is also relatively straightforward to implement these functionals in standard electronic structure codes (for checking purposes, see the absolute energies of three systems in Table S22 in the Supporting Information) and to derive the corresponding analytical derivatives. A comparison of the two new proposals finally shows that PWPB95-D3 clearly outperforms PTPSS-D3 in four out of five different statistical analyses. Many times, it has better MADs for the GMTKN30 set. The WTMADs are significantly lower, particularly at the triple-ζ level. The PTPSS error for the dissociation enthalpies of transition metal carbonyls is almost halved by PWPB95. Furthermore, PWPB95-D3 is mathematically less complicated and depends on fewer parameters. Moreover, only PWPB95-D3 can compete with the recently proposed DSD-BLYP functional. It is expected to also be useful for cases in which LYP correlation is known to fail. We therefore suggest general usage of PWPB95-D3 for future applications. The GMTKN30 database proved to be a useful tool for evaluating DFT methods, and further investigations along these lines are currently undertaken in our laboratories.

Acknowledgment. This work was supported by the "Fonds der Chemischen Industrie" with a scholarship to L.G. and by the Deutsche Forschungsgemeinschaft in the framework of the SFB 858 ("Synergetische Effekte in der Chemie—Von der Additivität zur Kooperativität"). We thank C. Mück-Lichtenfeld for his technical assistance.

Supporting Information Available: The formula for calulating the weighted total mean absolute deviation (WTMAD). Information about determining the s_6 for various MP2 versions and the double hydrids (Table S1). Information about the parameters fit sets (Tables S2 and S3). Results for GMTKN30 with (aug-)def2-QZVP (Tables S4−S11). Results for GMTKN30 with (aug-)def2-TZVPP (Tables S12−S19). Results for dissociation reactions of transition metal carbonyls (Table S20). Results for the study of HF and KS orbitals for five different molecules (Table S21). Reference values of PTPSS and PWPB95 for checking purposes, in case the functionals are implemented into other electronic structure codes (Table S22). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864–871.

(2) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(3) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: Oxford, U. K., 1989.

(4) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: New York, 2001.

(5) Dreizler, J.; Gross, E. K. U. *Density Functional Theory, An Approach to the Quantum Many-Body Problem*; Springer: Berlin, 1990.

(6) Rappoport, D.; Crawford, N. R. M.; Furche, F.; Burke, K. Approximate Density Functionals: Which Should I Choose? In *Computational Inorganic and Bioinorganic Chemistry*; Solomon, E. I., Scott, R. A., King, R. B., Eds.; Wiley-VCH: New York, 2009; pp 159−172.

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

(8) Görling, A.; Levy, M. *Phys. Rev. B* **1993**, *47*, 13105–13113.

(9) Görling, A.; Levy, M. *Phys. Rev. A* **1994**, *50*, 196–204.

(10) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786–4791.

(11) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

(12) Ángyán, J. G.; Gerber, I. C.; Savin, A.; Toulouse, J. *Phys. Rev. A* **2005**, *72*, 012510.

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

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

(15) Karton, A.; Tarnopolsky, A.; Lamere, J. F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.

(16) Sancho-García, J. C.; Pérez-Jiménez, A. J. *J. Chem. Phys.* **2009**, *131*, 084108.

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

(18) Graham, D. C.; Menon, A. S.; Goerigk, L.; Grimme, S.; Radom, L. *J. Phys. Chem. A* **2009**, *113*, 9861–9873.

(19) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2009**, *131*, 174105.

(20) Kozuch, S.; Gruzman, D.; Martin, J. M. L. *J. Phys. Chem. C* **2010**, *114*, 20801−20808.

(21) Zhang, Y.; Xu, X.; Goddard, W. A., III. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4963–4968.

(22) Zhang, I. Y.; Luo, Y.; Xu, X. *J. Chem. Phys.* **2010**, *132*, 194105.

(23) Zhang, I. Y.; Luo, Y.; Xu, X. *J. Chem. Phys.* **2010**, *133*, 104105.

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

(25) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys.* **2007**, *126*, 124115.

(26) Grimme, S.; Mück-Lichtenfeld, C.; Würthwein, E.-U.; Ehlers, A. W.; Goumans, T. P. M.; Lammertsma, K. *J. Phys. Chem. A* **2006**, *110*, 2583–2586.

(27) Grimme, S.; Antony, J.; Schwabe, T.; Mück-Lichtenfeld, C. *Org. Biomol. Chem.* **2007**, *5*, 741–758.

(28) Grimme, S.; Schwabe, T. *Acc. Chem. Res.* **2008**, *41*, 569–579.

(29) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2722–2729.

(30) Schwabe, T.; Grimme, S. *Eur. J. Org. Chem.* **2008**, 5928–5935.

(31) Schwabe, T.; Grimme, S. *J. Phys. Chem. A* **2009**, *113*, 3005–3008.

(32) Korth, M.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *5*, 993–1003.

(33) Schwabe, T.; Grimme, S. *J. Phys. Chem. Lett.* **2010**, *1*, 1201–1204.

(34) Grimme, S.; Djukic, J.-P. *Inorg. Chem.* **2010**, *49*, 2911–2919.

(35) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2010**, *6*, 107–126.

(36) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.

(37) Goerigk, L.; Grimme, S. *J. Phys. Chem. A* **2009**, *113*, 767–776.

(38) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611–4620.

(39) Goerigk, L.; Grimme, S. *J. Chem. Phys.* **2010**, *132*, 184103.

(40) Vintonyak, V. V.; Warburg, K.; Kruse, H.; Grimme, S.; Hübel, K.; Rauh, D.; Waldmann, H. *Angew. Chem., Int. Ed.* **2010**, *49*, 5902–5905.

(41) Huenerbein, R.; Schirmer, B.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6940–6948.

(42) Krieg, H.; Grimme, S. *Mol. Phys.* **2010**, *108*, 2655–2666.

(43) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.

(44) Grimme, S.; Kruse, H.; Goerigk, L.; Erker, G. *Angew. Chem., Int. Ed.* **2010**, *49*, 1402–1405.

(45) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.

(46) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095–9102.

(47) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793–9802.

(48) Almlöf, J. *Chem. Phys. Lett.* **1991**, *181*, 319–320.

(49) Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9677–9686.

(50) Zhang, I. Y.; Wu, J.; Xu, X. *Chem. Commun.* **2010**, *46*, 3057–3070.

(51) Prof. Stefan Grimme Research Web Site. http://www.uni-muenster.de/Chemie.oc/grimme/en/index.html (accessed December 2010).

(52) Gilbert, T. M. *J. Phys. Chem. A.* **2004**, *108*, 2550–2554.

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

(54) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(55) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460–4464.

(56) Grimme, S. *RICC: A coupled-cluster program using the RI approximation*; University of Münster: Münster, Germany, 2007.

(57) Pitonák, M.; Neogrády, P.; Cerný, J.; Grimme, S.; Hobza, P. *Chem. Phys. Chem.* **2009**, *10*, 282–289.

(58) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(59) Podeszwa, R.; Patkowski, K.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5974–5979.

(60) Grimme, S.; Steinmetz, M.; Korth, M. *J. Org. Chem.* **2007**, *72*, 2118–2126.

(61) Steinmann, S. N.; Csonka, G.; Carminboeuf, C. *J. Chem. Theory Comput.* **2009**, *5*, 2950–2958.

(62) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.

(63) Radom, L.; Hehre, W. J.; Pople, J. A. *J. Am. Chem. Soc.* **1971**, *93*, 289–300.

(64) Wodrich, M. D.; Jana, D. F.; von Ragué; Schleyer, P.; Corminboeuf, C. *J. Phys. Chem. A* **2008**, *112*, 11495–11500.

(65) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. *J. Chem. Phys.* **2006**, *124*, 114304.

(66) Goll, E.; Werner, H.-J.; Stoll, H. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3917–3923.

(67) Ogilvie, J. F.; Wang, F. J. H. *J. Mol. Struct.* **1992**, *273*, 277–290.

(68) Ogilvie, J. F.; Wang, F. J. H. *J. Mol. Struct.* **1993**, *291*, 313–322.

(69) Runeberg, N.; Pyykö, P. *Int. J. Quantum Chem.* **1998**, *66*, 131–140.

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

(71) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. *J. Chem. Phys.* **2005**, *123*, 62201.

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

(73) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

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

(75) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.

(76) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(77) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

(78) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(79) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(80) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklaβ, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. MOLPRO, version 2009.1. See http://www.molpro.net (accessed November 9, 2010).

(81) Kabelác, M.; Valdes, H.; Sherer, E. C.; Cramer, C. J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5000–5008.

(82) Tkatchenko, A., Jr.; , R. A. D.; Head-Gordon, M.; Scheffler, M. *J. Chem. Phys.* **2009**, *131*, 094106.

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

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

(85) Vázquez-Mayagoitia, A.; Sherrill, C. D.; Aprá, E.; Sumpter, B. G. *J. Chem. Theory Comput.* **2010**, *6*, 727–734.

(86) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(87) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544–2547.

(88) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(89) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Csonka, G. I.; Scuseria, G. E. *Phys. Rev. A* **2007**, *76*, 042506.

(90) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. *Phys. Rev. Lett.* **2009**, *103*, 026403.

(91) Perdew, J. P. In *Proceedings of the 21st Annual International Symposium on the Electronic Structure of Solids*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.

(92) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040–1046.

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

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

(95) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

(96) TURBOMOLE: Ahlrichs, R. et al. Universität Karlsruhe: Karlsruhe, Germany, 2008. See http://www.turbomole.com. (accessed November 9, 2010).

(97) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(98) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(99) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(100) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(101) Metz, B.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2000**, *113*, 2563–2569.

(102) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113–11123.

(103) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

(104) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(105) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

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

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

(108) Hyla-Kryspin, I.; Grimme, S. *Organometallics* **2004**, *23*, 5581–5592.

(109) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(110) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(111) Kind, C.; Reiher, M.; Neugebauer, J.; Hess, B. A. *SNF*, version 2.2.1; Universität Erlangen: Erlangen, Germany, 2002.

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

(113) Stowasser, R.; Hoffmann, R. *J. Am. Chem. Soc.* **1999**, *121*, 3414–3420.

(114) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.

(115) Teale, A. M.; Coriani, S.; Helgaker, T. *J. Chem. Phys.* **2010**, *132*, 164115.

(116) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221–7230.

(117) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014–6029.

(118) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 10478–10486.

(119) Guner, V.; Khuong, K. S.; Leach, A. G.; Lee, P. S.; Bartberger, M. D.; Houk, K. N. *J. Phys. Chem. A* **2003**, *107*, 11445–11459.

(120) Ess, D. H.; Houk, K. N. *J. Phys. Chem. A* **2005**, *109*, 9542–9553.

(121) Dinadayalane, T. C.; Vijaya, R.; Smitha, A.; Sastry, G. N. *J. Phys. Chem. A* **2002**, *106*, 1627–1633.

(122) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715–2719.

(123) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.

(124) Neese, F.; Schwabe, T.; Kossmann, S.; Schirmer, B.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *5*, 3060–3073.

(125) Zhao, Y.; Tishchenko, O.; Gour, J. R.; Li, W.; Lutz, J. J.; Piecuch, P.; Truhlar, D. *J. Phys. Chem. A* **2009**, *113*, 5786–5799.

(126) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(127) Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 204112.

(128) Piacenza, M.; Grimme, S. *J. Comput. Chem.* **2004**, *25*, 83–99.

(129) Woodcock, H. L.; Schaefer, H. F., III; Schreiner, P. R. *J. Phys. Chem. A* **2002**, *106*, 11923–11931.

(130) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635–3638.

(131) Lepetit, C.; Chermette, H.; Gicquel, M.; Heully, J.-L.; Chauvin, R. *J. Phys. Chem. A* **2007**, *111*, 136–149.

(132) Lee, J. S. *J. Phys. Chem. A* **2005**, *109*, 11927–11932.

(133) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A., III. *J. Chem. Theory Comput.* **2009**, *5*, 1016–1026.

(134) Reha, D.; Valdes, H.; Vondrasek, J.; Hobza, P.; Abu-Riziq, A.; Crews, B.; de Vries, M. S. *Chem.—Eur. J.* **2005**, *11*, 6803–6817.

(135) Gruzman, D.; Karton, A.; Martin, J. M. L. *J. Phys. Chem. A* **2009**, *113*, 11974–11983.

(136) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. *J. Chem. Theory Comput.* **2009**, *5*, 679–692.

(137) Wilke, J. J.; Lind, M. C.; Schaefer, H. F., III; Császár, A. G.; Allen, W. D. *J. Chem. Theory Comput.* **2009**, *5*, 1511–1523.

CT100466K

JCTC Journal of Chemical Theory and Computation

A CASPT2 Description of the Electronic Structures of $FeO_3^{-/0}$ in Relevance to the Anion Photoelectron Spectrum

Van Tan Tran and Marc F. A. Hendrickx*

Afdeling Kwantumchemie en Fysicochemie, Departement Chemie, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001 Heverlee-Leuven, Belgium

Received September 15, 2010

Abstract: DFT and multireference methods were used to investigate the electronic structure of FeO_3 and FeO_3^- clusters. Geometries of different spin multiplicities and conformations were optimized without any symmetry restrictions at the BP/QZVP level and further refined with the CASPT2 method. Although the latter type of calculations were performed by using the C_{2v} point group, all low-lying states relevant to the photoelectron spectrum were found to correspond to or to resemble closely a planar D_{3h} iron trioxide with no bonds between the oxygen atoms. Depending on the computational method used, the ground state of the FeO_3^- anion can be either $^2E''$ or $^4A_1'$. The two lowest binding energy bands of the photoelectron spectrum of FeO_3^- can only be ascribed to electron detachments from the $^2E''$ state. The first band is the result of a transition to the $^1A_1'$ ground state of FeO_3, whereas the second band originates from the first excited $^3E''$ state. A harmonic vibrational analysis of the symmetric stretch shows that the observed vibrational progressions of these two bands in the photoelectron spectrum of FeO_3^- are also in line with the assignment. A molecular orbital analysis led to the conclusion that the electronic structures of the anionic and neutral clusters can formally be described by an oxidation state of iron of +5 and +6, respectively. A population analysis, on the contrary, points to an ionization that takes place on the oxygen atoms.

Introduction

Iron oxide clusters FeO_n can serve as basic models for numerous biological systems[1,2] that play important roles, such as oxygen transportation. In an industrial context, they are also of importance because they bear relevance to the main factor in redox processes like iron corrosion[3] and as a model for various catalysts.[4−8] To understand the electronic structures of FeO_n/FeO_n^-, Wang and co-workers synthesized FeO_n^- ($n = 1-4$) in the gas phase and investigated their photoelectron spectroscopy (PES) experimentally by using lasers with photon energies of 3.49 and 4.66 eV.[9−12] More than a decade ago, Chertihin and co-workers also observed the neutral systems of these clusters by infrared spectroscopy as products of the reaction of laser-ablated Fe atoms with oxygen molecules in a condensing argon stream.[13] By analyzing their infrared spectra in detail and with the aid of calculations, the authors proposed and confirmed the existence for various stable conformations of iron−oxygen clusters. Their theoretical work comprised geometry optimizations and harmonic frequency calculations at the BP/DZVP level. Also, results obtained by the B3LYP hybrid technique with a (15s12p6d1f)/[9s7p4d1f] basis set for iron and 6-311+G(2df) basis set for oxygen were given. A few years later, density functional theory (DFT) calculations employing the 6-311+G* basis sets and the BPW91 functional were performed by Gutsev et al.,[14] mainly with the purpose of calculating the electron affinities of FeO_n. Rather recently, a combined experimental and theoretical study on the neutral FeO_3 was carried out by Yu et al.[15] by using matrix infrared spectroscopy combined with DFT calculations. To date, all experimental and computational works agree on a stable η^2-O_2FeO complex that is characterized by a quintet lowest state of planar C_{2v} symmetry and a more

* Corresponding author. E-mail: marc.hendrickx@chem.kuleuven.be.

10.1021/ct1005246
Published on Web 01/06/2011

stable iron trioxide possessing a planar singlet ground state of D_{3h} symmetry. A detailed description of the electronic structure of the monoiron oxide molecules is however still needed.

Unsaturated complexes of the type at hand are known to possess many low-lying states of different spin multiplicities.[14,16−24] As a consequence, for the identification of their ground states and all possible low-lying excited states, a multireference wave function method turns out to be a very convenient tool. Indeed, these methods, such as CASPT2 (complete active space second-order perturbation theory) or MRCI (multireference configuration interaction), were applied successfully for the description of the electronic structures of FeO/FeO^- [25,26] and FeO_2/FeO_2^-.[27] It is very surprising that the electronic structures of the higher monoiron oxides clusters, such as $FeO_3^{-/0}$, are still not studied at any of these computational levels and, therefore, remain a bit of a mystery. More specifically, the additional electron was found to be delocalized over the three oxygen atoms,[13] whereas in recent CASPT2 studies it was argued that these clusters can be described as transition metal complexes, implying that the valence orbitals have predominantly 3d metal character.[17,19,25] In this contribution, we present a multireference wave function study on the basis of the CASPT2 method for FeO_3 and FeO_3^-, containing a detailed picture for their electronic structures. The accuracy of the presented results are tested against the experimental photoelectron spectrum of FeO_3^-.

Computational Details

In the first step of our study, the geometries of different spin multiplicities from singlet to sextet of FeO_3 and FeO_3^- are optimized by an unrestricted DFT technique with the QZVP basis sets[28] and the BP86 functional[29] as implemented in the TURBOMOLE 5.10 suite of programs.[30] The geometry optimizations were done without symmetry restrictions so that all structural parameters are allowed to relax and the lowest energy electronic states of either of the two molecules are serious candidates for their actual ground states. At all of these optimized geometries, a harmonic vibrational frequency calculation is carried out to ensure that the structures found are true minima and not saddle points.

In the second step, we utilized the efficient MOLCAS 7.4 package[31] to perform CASPT2 calculations. Since only Abelain point groups are implemented in this computational package, all of the calculations were carried out using C_{2v} symmetry, although some optimized DFT geometries exhibit a higher D_{3h} symmetry. As illustrated in Figure 1, the molecule is placed in the *yz* plane with one oxygen atom along the *z* axis. The basis sets used are [8s,7p,5d,4f,2g] and [6s,5p,4d,2f] ANO-RCC, for iron and oxygen, respectively.[32] Scalar relativistic effects are included by the Douglas−Kroll formalism.[33,34] The orbitals for the CASPT2 calculations are obtained from complete active space self-consistent field (CASSCF) calculations, for which the active spaces include the 2p orbitals of oxygen and the 3d and 4s orbitals of iron. Corresponding to either the neutral or anionic system, we have a total of 20 or 21 electrons in 15 orbitals. The

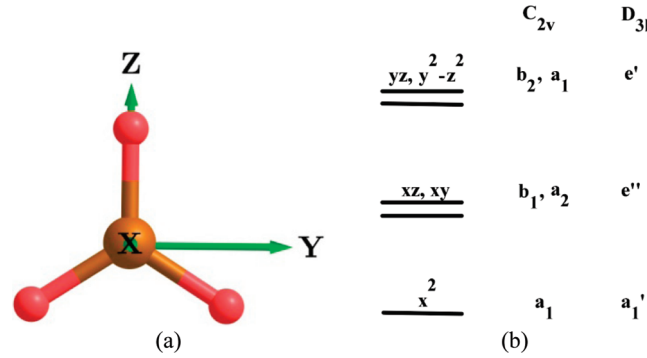


Figure 1. (a) Choice of the coordination system for the CASPT2 calculations. (b) Qualitative orbital energy scheme for the valence d orbitals.

CASPT2 calculations correlated the 3p, 3d, and 4s orbitals of iron and 2s and 2p orbitals of oxygen, whereas at this level the intruder states problem was addressed by applying an imaginary shift. The BP/QZVP structures were refined using numerical gradient CASPT2 optimization. From these structures, adiabatic detachment energies (ADEs) and vertical detachment energies (VDEs) for FeO_3^- were evaluated. In order to improve our results, single point calculations with the larger ANO-RCC basis set [8s,7p,6d,4f,2 g,1 h] for iron and [7s,6p,4d,3f,1 g] for oxygen were performed.

Results and Discussion

Both previous DFT studies[14,15] showed that the FeO_3 systems can have two relatively low-lying conformations. One that does not possess any O−O bonds is denoted as O_3Fe, whereas another has one O−O bond and corresponds to η^2-O_2FeO. Linear structures of the type O−Fe−O−O were calculated at energies 50 kcal/mol higher than the ground state.[13] With the purpose of investigating the photoelectron spectrum of FeO_3^-, we therefore do not need to consider them in our study. In order to be sure about the low-lying states of both anion and neutral systems, we reoptimized various spin multiplicities of O_3Fe and η^2-O_2FeO at the BP/QZVP level. The results are summarized in Figure 2, which contains the structural parameters for the various conformations and their relative energies. For the *O_3Fe conformation*, the ground state of the neutral system clearly turns out to be a singlet state with D_{3h} symmetry, while the ground state of the anion can be either a doublet possessing C_{2v} symmetry or a quartet with D_{3h} symmetry. Indeed, the quartet is predicted at this computational level to be just 0.02 eV higher in energy than the doublet. Our DFT calculations predict a value 1023 cm^{-1} for the antisymmetric stretching modes of the singlet ground state, see Table 1, which is in fairly good agreement with the experimental values of 975.8 cm^{-1} and 950 cm^{-1} obtained from the infrared spectrum in a argon matrix.[13,15] For the *η^2-O_2FeO conformation* and in accordance with previous theoretical work, we could identify a quintet of C_s symmetry and a quartet of C_{2v} symmetry as the lowest states of the neutral and anionic systems, respectively. Our predicted vibrational frequencies of 933 cm^{-1} and 1052 cm^{-1} for this quintet are in good agreement with the infrared bands at 1148 cm^{-1} and 928 cm^{-1} of ref

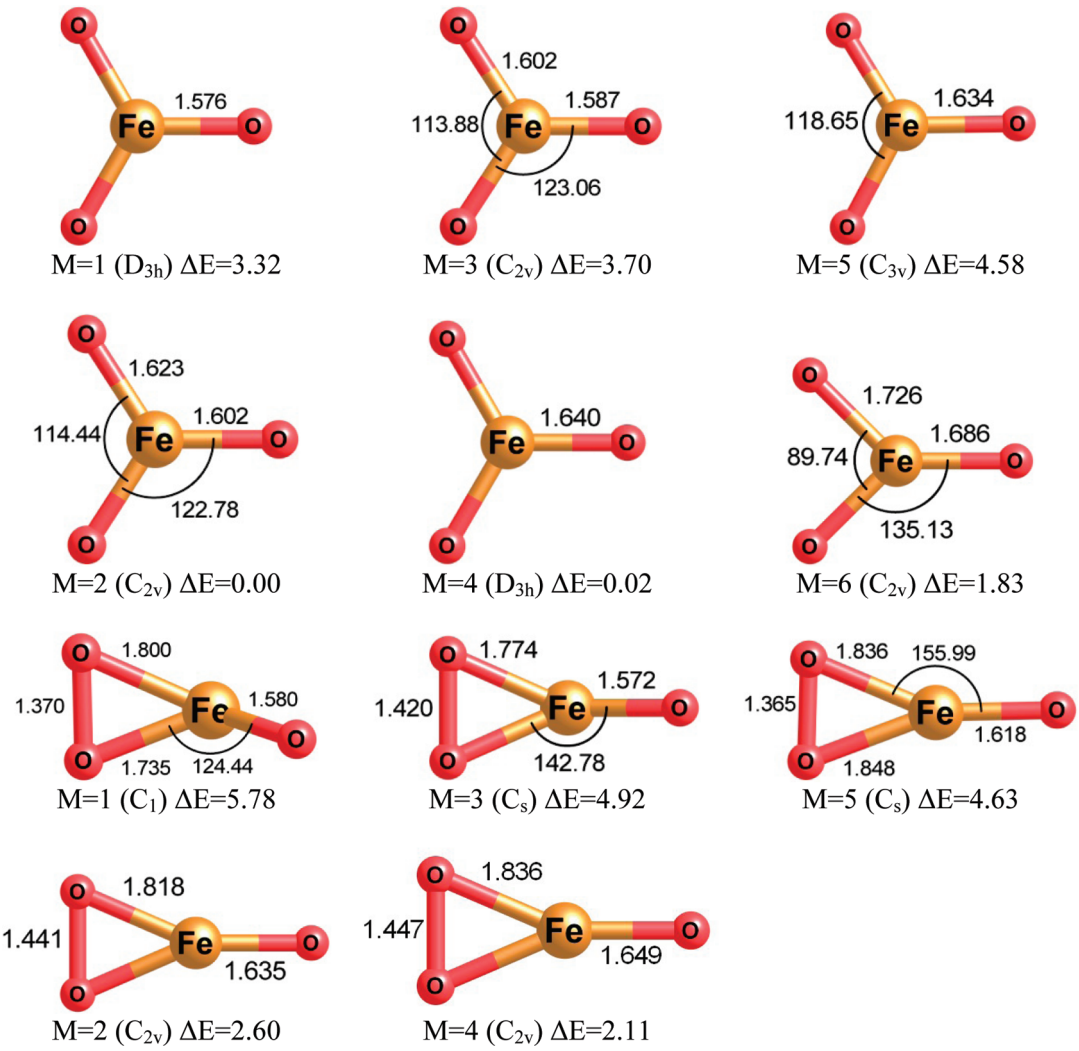


Figure 2. Structures (bond distances in Ångstroms and bond angles in degrees) and relative energies (eV) of FeO_3 and FeO_3^- as obtained by BP/QZVP calculations.

13 and the corresponding 1002 cm^{-1} band of ref 15 as well as with the DFT values mentioned in both papers. The predicted quintet lowest energy state of η^2-O_2FeO is, energetically, 1.3 eV higher than the singlet ground state of the O_3Fe conformation. For the anionic conformations, the quartet of η^2-O_2FeO^- is 2.11 eV higher than the doublet state of O_3Fe, which was calculated as the ground state of the FeO_3^- anion. Because of this large energy difference, it is safe to base our assignment of the photoelectron spectroscopy of FeO_3^- only on the O_3Fe conformation. With this information at hand, we only investigated the $O_3Fe^{-/0}$ conformations in our subsequent CASPT2 calculations, which we will simply denote as $FeO_3^{-/0}$.

Starting from the optimized geometries of the previous DFT calculations, we first performed for each specific spin multiplicity CASPT2 single point calculations for the purpose of finding the lowest energy state for each irreducible representation of the C_{2v} point group. In a next step, all possible low lying states were reoptimized at the CASPT2 level by employing the small ANO-RCC basis set as mentioned in the Computational Details. In Figure 3, we present the optimized structures of the $^1A_1$, $^3A_2$, $^3B_1$, $^2A_2$, $^2B_1$, and $^4B_2$ states. The structural parameters of these states were found to be nearly unchanged from the DFT geometries of the same multiplicity. To a certain extent, this confirms that we have found the lowest energy states. For the neutral FeO_3, there can be little doubt about its ground state. This is clearly the $^1A_1$, while at 0.34 eV higher in energy we calculated the first excited state as $^3A_2$. This result is more or less coherent with the DFT calculations. For FeO_3^-, there is a contradiction between CASPT2 and DFT concerning the prediction of the ground state. The $^4B_2$ state is now 0.27 eV lower than the $^2A_2$ state. However, this energy difference is only just outside the believed error margin of the CASPT2 method. Furthermore, by taking into account that at the employed computational level it is likely to overestimate the relative stability of the higher spin state, we cannot make an absolute statement about the true ground state of FeO_3^-. We believe that a more secure statement about the ground state of FeO_3^- most likely needs a more accurate treatment of the electronic dynamic correlation energy.

The CASPT2 geometry optimization shows that the symmetry of $^4B_2$ and $^1A_1$ is actually D_{3h} while the symmetry of $^2A_2$ and $^3A_2$ is C_{2v}. These C_{2v} structures exhibit only a slight distortion from D_{3h} symmetry as a result of weak Jahn−Teller effects. From single point CASPT2 calculations,

Table 1. Relative Energies (RE), Harmonic Unscaled Vibrational Frequencies (cm^{-1}), and Intensities (km/mol) for the Two Studied Conformations of FeO_3 and FeO_3^- at the BP/QZVP Level

cluster	spin multiplicity	RE (eV)	frequency (intensity)
O_3Fe^-	2	0.00	195 (50), 237 (3), 335 (0), 817 (407), 856 (0), 969 (197)
	4	0.02	178 (52), 310 (0), 311 (0), 819 (0), 913 (204), 914 (204)
	6	1.83	120 (7), 143 (47), 217 (0), 689 (4), 732 (6), 811 (148)
O_3Fe	1	3.33	147 (11), 332 (0), 333 (0), 920 (0), 1023 (101), 1023 (101)
	3	3.70	99 (41), 200 (22), 345 (0), 864 (1), 865 (175), 995 (92)
	5	4.58	80 (33), 219 (0) 223 (0), 605 (2), 606 (2), 837 (2)
η^2-O_2FeO^-	2	2.60	86 (1), 126 (1), 438 (7), 561 (6), 882 (13), 936 (288)
	4	2.11	130 (11), 165 (9), 537 (6), 549 (10), 869 (30), 918 (317)
η^2-O_2FeO	1	5.78	158 (7), 234 (18), 526 (3), 637 (1), 967 (101), 1023 (166)
	3	4.92	122 (19), 187 (16), 394 (16), 571 (4), 938 (15), 1002 (200)
	5	4.63	44 (15), 146 (24), 506 (1), 611 (2), 933 (16), 1052 (174)

as given in Tables 2 and 4, we know that $^4B_2$ and $^1A_1$ are nondegenerate states, while $^2A_2$ and $^2B_1$ and $^3A_2$ and $^3B_1$ are degenerate D_{3h} states. In both cases, they are the two components of E″ states. A graphical representation for the potential energy curves of the low-lying states is given in Figure 4. Here, a geometry optimization for the low lying states within D_{3h} symmetry is performed by carrying out CASPT2 single point calculations for a number of Fe−O bond distances, around the respective equilibrium bond distances. Although the calculations were carried out within the broken symmetry approach, the A_2 and B_1 curves are shown to be nearly degenerated; the scale used on the ordinate axis places the states at the same points on the two graphs of the $^2E''$ and $^3E''$ states. The energetic Jahn−Teller effects on both states were further examined by performing a CASPT2 geometry optimization for their B_1 components, a procedure that takes full account of these effects as applied previously on the related $FeO_4^{-/0}$ clusters at the DFT level.[35,36] The results are given Table 2 and are depicted in Figure 3 and show only small changes in geometry. Indeed, the bond distances between two E″ components vary about 1/100 of an Ångstrom. Bond angles between the A_2 and B_1 components differ only by a few degrees. There is here however a systematic trend that can be observed. The B_1 component has two bond angles smaller than the D_{3h} value of 120° and consequently a third bond angle that has increased from this high symmetry value. The opposite is found for the A_2 components. As a result of having two larger bond angles, the remaining angle decreases noticeably further from 120° in these components. The equilibrium distances of the bonds that are involved in the smaller bond angles are partly larger due to the increased electrostatic repulsion between the negatively charged oxygens. A further explanation for these geometric differences will be discussed after a detailed molecular orbital analysis of the electronic structure of the states in the following paragraph. As Jahn−Teller effects decrease the energy, the A_2 states are stabilized more since they possess the largest distortions. This causes $^2A_2$ and $^3A_2$ to become the lowest excited CASPT2 states of FeO_3^- and FeO_3, respectively. Energetically, the effects are indeed very small. The difference between the two components is calculated to be on the order of a few hundredths of an electronvolt in Table 2. These values are by far too small to induce any effect on the experimental photoelectron spectra or to have a measurable effect on the electron affinity.

As mentioned in the previous paragraph, there is a need for an in-depth description of the electronic structure of FeO_3^-. In view of the intended assignment of the photoelectron spectrum, this is best done in relationship with its neutral counterpart. For doing so, we made plots of the molecular orbitals of the active spaces for all relevant states. These plots for the $^4B_2$ CASPT2 ground state, classified according to the irreducible representation of the C_{2v} point group and accompanied by their CASSCF natural occupation numbers, are depicted in Figure 5. A comparison with similar orbitals obtained with the CASSCF calculations on all mentioned states shows that the valence orbitals of the studied molecules, i.e., the orbitals with occupation numbers in the vicinity of one, have always a predominant iron 3d character. Further observation shows that they are of an antibonding nature between iron and the adjacent oxygen atoms. From the entry for the leading configuration of $^4B_2$ in Table 4, we learn that the three unpaired electrons of this quartet state are indeed occupying the iron 3d type orbitals. More specifically, one electron occupies a nearly pure d_{x^2} orbital in the $13a_1$ orbital. A second unpaired electron can be found in the $5b_1$ orbital with predominant d_{xz} character, while $2a_2$ as a similar d_{xy} orbital hosts the third unpaired electron. For D_{3h} geometries, d_{x^2} is totally symmetric (a_1'), whereas the latter two d orbitals form a basis for the e″ irreducible representation. The remaining two d orbitals are depicted as $14a_1$ ($d_{y^2-z^2}$) and $8b_2$ (d_{yz}). They are not occupied in any of the states relevant to the photoelectron spectrum because they are positioned at much higher energies. These d orbitals belong to an antibonding 2-fold degenerate e′ level. Molecular orbitals with predominant oxygen contributions are without any exception doubly occupied. Pure σ binding interaction between iron and oxygen is present in the $10a_1$, $12a_1$, and $5b_2$ orbitals. The other predominant oxygen orbitals are either pure π bonding orbitals, i.e., $3b_1$, $4b_1$, $7b_2$, and $1a_2$, or a mixture of the two types of bonding: $11a_1$ and $6b_2$. All of these orbitals are to an extremely large extent constructed from the 2p orbitals of the oxygen atoms. Since the 2s orbitals of these atoms are all in the inactive space, the electronic structure picture emerging from the above orbitals puts a formal charge of +5 on iron and −2 on each oxygen center. All of the other low-lying states of the anionic cluster were found to agree with this picture. A similar orbital analysis for the relevant FeO_3 states agrees completely with the above orbital picture for FeO_3^-, which therefore renders a formal oxidation state of +6 for the metal atom in the neutral cluster. For the neutral iron trioxide cluster, the same conclusion regarding the oxidation state of iron was put forward by Gong and Zhou after analyzing their DFT

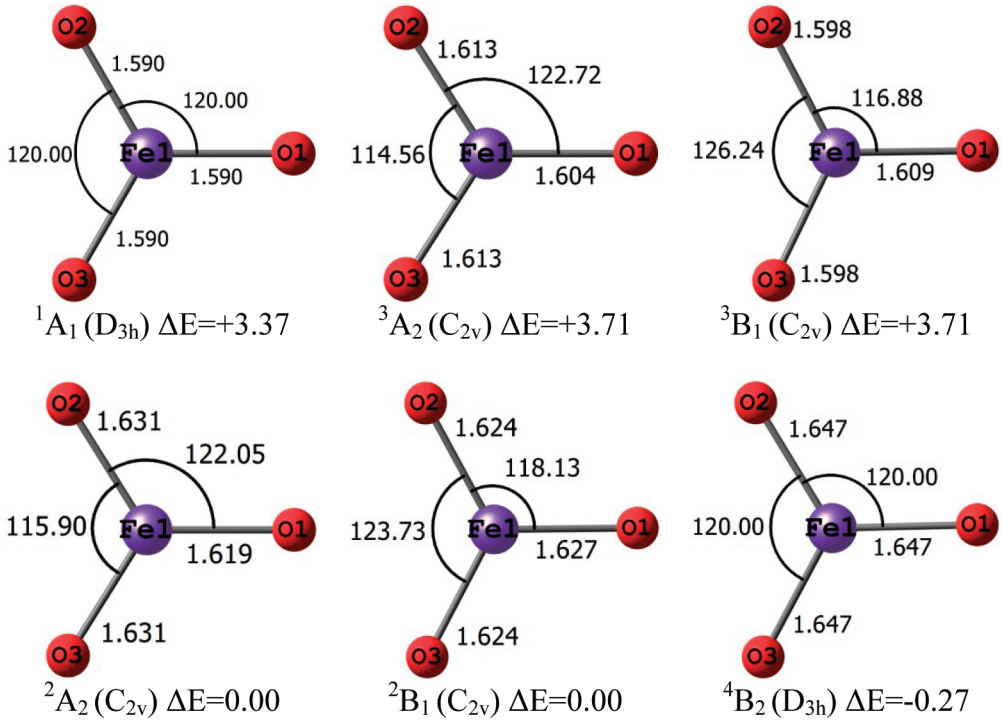


Figure 3. Structures (bond distances in Ångstroms and bond angles in degrees) and relative energies (eV) for the low lying states of FeO_3 and FeO_3^- as obtained by CASPT2 geometry optimizations.

Table 2. Relative CASPT2 Energies for the Iron Trioxides FeO_3 and FeO_3^-

		relative energy (eV)		
cluster	state	*a*	*b*	*c*
FeO_3^-	$^2A_2$	0.00	0.00	0.00
	$^2B_1$		0.00	0.02
	$^4B_2$	−0.26	−0.27	−0.25
FeO_3	$^1A_1$	3.40	3.37	3.44
	$^3A_2$	3.74	3.71	3.77
	$^3B_1$		3.71	3.77

[a] ASPT2 single point calculations with small ANO-RCC basis sets at BP/QZVP geometries. [b] CASPT2 geometry optimizations with small ANO-RCC basis sets. [c] CASPT2 single point calculation with large ANO-RCC basis sets at geometries b.

Table 3. Mulliken Population Analysis Charges for Low-Lying States of the FeO_3 and FeO_3^- Clusters As Obtained from the CASPT2 Wave Functions

	Mulliken charge (e^-)			
state	Fe	O(1)	O(2)	O(3)
$^2A_2$	+0.91	−0.59	−0.66	−0.66
$^2B_1$	+0.97	−0.70	−0.63	−0.63
$^4B_2$	+1.13	−0.71	−0.71	−0.71
$^1A_1$	+1.10	−0.37	−0.37	−0.37
$^3A_2$	+1.24	−0.40	−0.42	−0.42
$^3B_1$	+1.29	−0.45	−0.42	−0.42

results.[15] Formally, the ionization processes that lie beneath the photoelectron spectrum of FeO_3^- involve the removal of an iron 3d electron, a result that is apparently in sharp contrast to the previous DFT study.[14] Compared to FeO_3 and on the basis of a natural bond analysis of the DFT orbitals, the extra electron of FeO_3^- was found by these authors to be delocalized over the oxygens. So, at first sight, a formal description of the electronic structure and an analysis of the

Table 4. Vertical Detachment Energies (VDE) from the $^4B_2$ State As Calculated by CASPT2 with Small ANO-RCC Basis Sets

Cluster	State	Leading Configuration CASSCF[a]	VDE (eV)
FeO_3^-	$^4A_1$	$11a_1^2 12a_1^2 13a_1^2 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^1$	1.74
	$^4A_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^1\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	3.14
	$^4B_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^1\ 2a_2^1$	1.69
	$^4B_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	0.00
FeO_3	$^3A_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^1$	4.93
	$^3A_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	4.06
	$^3B_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^0$	4.06
	$^3B_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^1\ 2a_2^0$	5.64
	$^5A_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^1$	4.82
	$^5A_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^1 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	5.47
	$^5B_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^1\ 4b_1^2 5b_1^0\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^1$	6.29
	$^5B_2$	$11a_1^2 12a_1^1 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	5.89
	1^3B_2	$11a_1^2 12a_1^2 13a_1^0 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	5.57
	2^3B_2	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^1\ 2a_2^0$	5.82

[a] In the leading configuration, the $10a_1$, $3b_1$, $5b_2$, and $1a_1$ orbitals are always doubly occupied; the $15a_1$ orbital is always unoccupied.

distribution of the electrons over the constituent atoms lead to opposing underlying electron detachment processes. An explanation for this discrepancy is given by a Mulliken

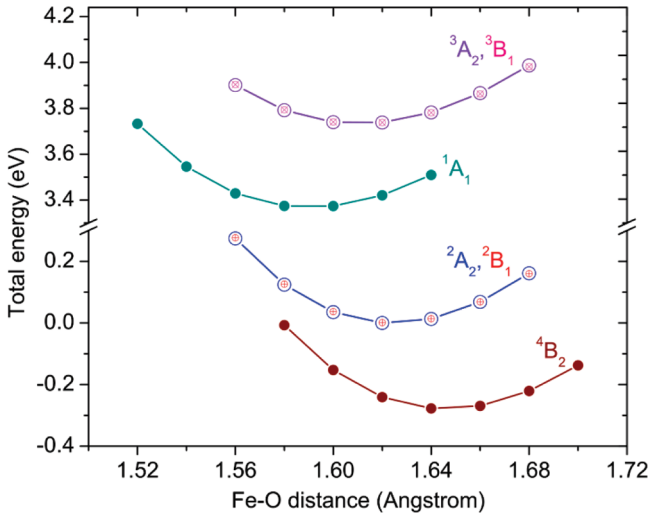


Figure 4. CASPT2 (using small ANO-RCC basis sets) potential energy curves of the symmetric Fe−O bond stretch (D_{3h} symmetry).

population analysis of the CASPT2 wave functions, which is given in Table 3. In agreement with the DFT description, the charge on the iron atom is hardly different between the anionic and neutral states. The formal removal of an electron for either three-occupied d orbital is compensated by small but effective relaxations in the underlying closed shell oxygen orbitals, which increase the σ and π ligand to metal donation of electronic charge. For a localization of charges within the molecules, a population analysis is therefore more trustworthy. The formal description of the electronic structure however has its merits in the sense that it provides a qualitative understanding of the relative position of the low-lying states of both species, as we will illustrate in the next paragraph.

The above description of the electronic structure closely resembles that of the classical and well-known transition metal complexes. For this type of compound, the electronic structure has, for decades, very successfully been described by qualitative ligand field theory. Within this model, the lower lying ligand (O^{2-} ions in present case) valence orbitals interact with the higher positioned d orbitals of the metal and become bonding molecular orbitals. For planar D_{3h} complexes, the ligand field model leads to the characteristic energy pattern in which the higher energy antibonding d orbitals are split into three levels (Figure 1). A lowest almost nonbonding orbital is only weakly σ antibonding, as illustrated for $13a_1$ in Figure 5. In the middle, we find an exclusively π antibonding e″ level ($5b_1$ and $2a_2$ orbitals in Figure 5). The highest d level corresponds to the $14a_1$ and $8b_2$ orbitals of the mainly σ antibonding e′ shell. For FeO_3^-, the remaining three metal valence electrons are to be distributed among the lowest d orbitals. For the anion, the first candidate ground state is found by placing two electrons in the lowest $13a_1$ orbital and the remaining electron in either one of the two d_π ($5b_1$ or $2a_2$) orbitals, giving rise to a so-called low-spin $^2E''$ state, which is Jahn−Teller active. The occupation of the antibonding d_{xy} ($2a_2$) orbital induces an increase in two O−Fe−O bond angles and an increase in two Fe−O bond distances, to which this orbital is oriented. This is the result of the increased electrostatic interactions between the approaching oxygen atoms and the antibonding nature of the $2a_2$ orbital. The entry for the $^2A_2$ of Figure 3 is in accordance with this analysis. The occupation, on the other hand, of d_{xz} ($5b_1$) causes one bond angle to increase, which stabilizes the $^2B_1$ component. Opening up two bonds is apparently more favorable, which makes the $^2A_2$ component the lowest in energy. The second candidate ground state is found by distributing the three valence metal electrons equally among the three lowest orbitals with the same spin. The resulting $^4B_2$ ($^4A_1'$ in D_{3h}) has a high-spin state with a lower electron repulsion and is predicted by CASPT2 as the ground state of FeO_3^-. For the neutral cluster, we expect two low lying states. The first one places two valence electrons in the lowest d_{x^2} orbital, giving a $^1A_1'$ state, which is always predicted as the ground state. The first excited state is a $^3E''$ with an occupation of $(d_{x^2})^1(d_{xy}, d_{yz})^1$. Quintets are expected at much higher energies because they involve an ionization process from one of the low-lying oxygen 2p orbitals. Thus, we suppose that they have no relevance to the low-energy part of the photoelectron spectrum. In this context, the above ligand field picture for these complexes attributes important roles to the $^1A_1'$ and the $^3E''$ states of the neutral complex.

The photoelectron spectrum of FeO_3^- measured with 4.66 eV laser photon energy, shown in Figure 6, has two low-lying bands.[9−12] The first band starts at 3.26 eV (X band), and the second band is located at higher binding energies (A band). The structure of the X band is composed of three sharp peaks of more or less equal intensity. Together, they are proposed to form a vibrational progression. The A band consists of two peaks, of which the first one at 3.81 eV has a much larger intensity than the higher energy peak. Both states of FeO_3 are measured to have a vibrational frequency of 850 cm^{-1}. These two simple vibrational structures indicate, according to the authors, that the geometries of FeO_3 and FeO_3^- possess a high level of symmetry like D_{3h}. Further, it was argued that the observed electron detachment processes induce only very slight geometry changes along the Fe−O stretching coordinate. On a qualitative basis, our DFT and CASPT2 geometry optimizations confirm these conclusions. Indeed, due to the extremely small Jahn−Teller effects, all of the states of the neutral and the anionic iron trioxide species that are relevant to the photoelectron either possess a D_{3h} structure (spacial nondegenerate states) or only slightly deviate from it (spacial degenerate states). In the latter case, the energy barriers that connect the two minima of a Jahn−Teller active state of D_{3h} symmetry are evaluated by CASPT2 as being very small. In regard to vibrational considerations, these states can be seen as having the higher D_{3h} symmetry, and therefore the assignment of the photoelectron spectrum should be carried out on these grounds. To answer the question of which state of FeO_3^- is responsible for its photoelectron spectrum, we calculated the one electron ADEs and VDEs from the $^2E''$ and $^4A_1'$ states.

According to the CASPT2 results, $^4A_1'$ has the lowest energy of FeO_3^-, so we need to explore the ionizations from this state first. Following the one-electron detachment selection rule, we only need to consider triplet and quintet

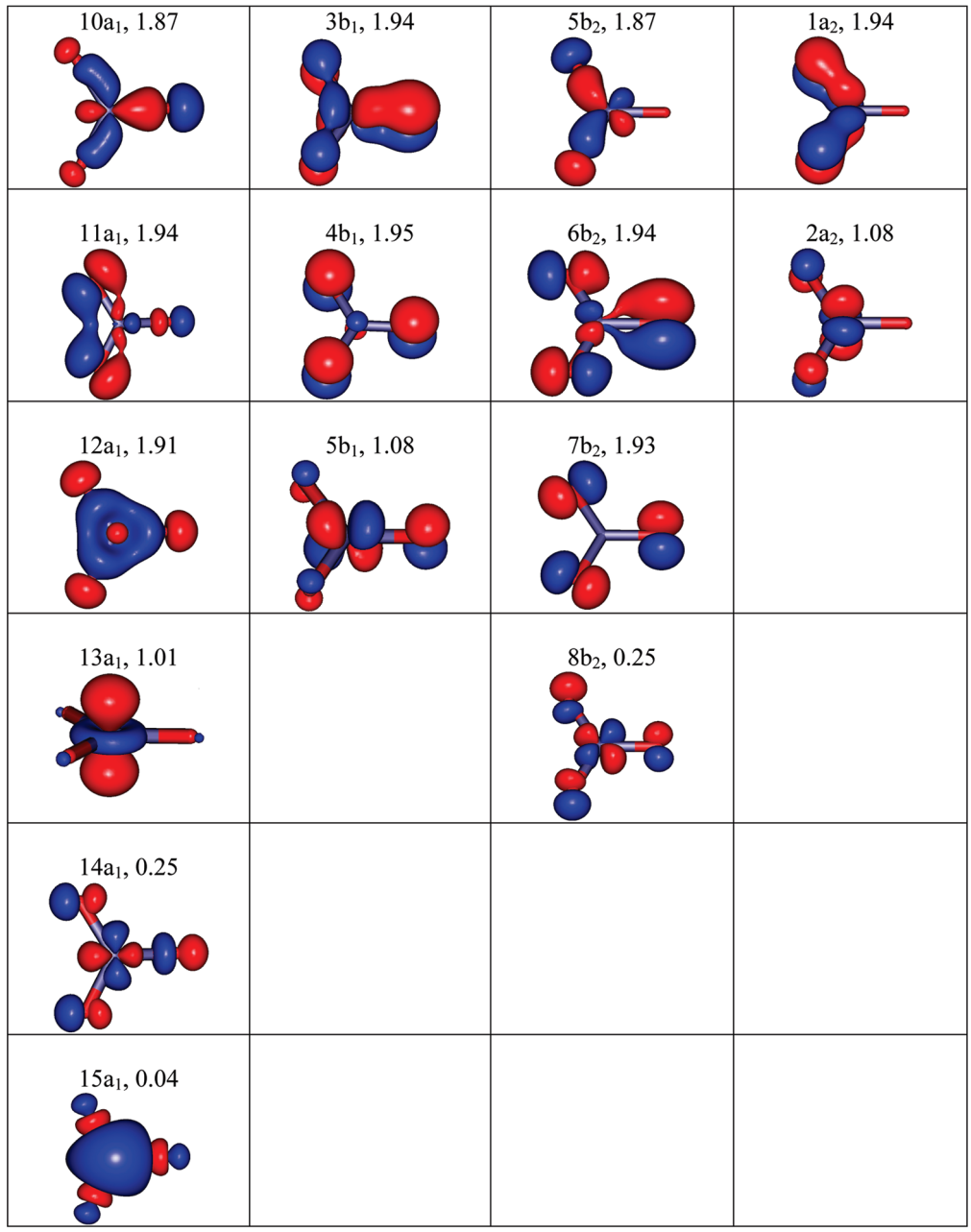


Figure 5. Pseudonatural molecular orbital plots and their occupation numbers for the $^4B_2$ state (FeO_3^-) as calculated by CASSCF (small ANO-RCC basis sets).

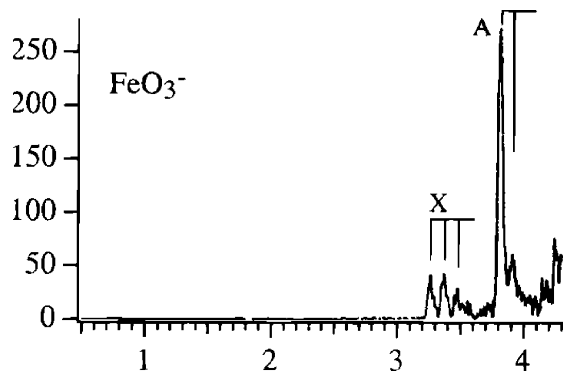


Figure 6. Photoelectron spectrum of FeO_3^- taken from ref 10. Abscissa: binding energies in electronvolts. Ordinate: relative electron intensities.

states. Table 4 summarizes the lowest VDEs for each irreducible representation of C_{2v} and relevant spin multiplicities. As a first conclusion, we note that the VDEs needed to reach the quintets are much higher than those for the triplets, a finding that already could be made from our DFT calculations and in agreement with the previous computational study.[14] Our results point out that removing one electron from the configuration of $^4A_1'$ (or $^4B_2$ in C_{2v}) can give only the following low-lying C_{2v} triplet states: $^3A_2$ and $^3B_1$. The removal of an electron from the $5b_1$ orbital of $^4B_2$ creates $^3A_2$, whereas an ionization from the $2a_2$ orbital results in the $^3B_1$ state of the neutral system. As a consequence of the small energetic Jahn−Teller effects, both transitions occur with a VDE value of 4.06 eV, and they could correspond to the A band in photoelectron spectrum of FeO_3^-. The quintet states $^5A_1$, $^5A_2$, and $^5B_2$ can also be formed by a one-electron

Table 5. Vertical Detachment Energies (VDE) from the $^2A_2$ State As Calculated by (a) CASPT2 with Small ANO-RCC Basis Sets and (b) CASPT2 with Large ANO-RCC Basis Sets[a]

cluster	state	leading configuration CASSCF	VDE (eV) a	VDE (eV) b	exp (eV)
FeO_3^-	$^2A_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^2$	1.03		
	$^2A_2$	$11a_1^2 12a_1^2 13a_1^2 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	0.00		
	$^2B_1$	$11a_1^2 12a_1^2 13a_1^2 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^0$	0.12		
	$^2B_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	0.75		
FeO_3	$^1A_1$	$11a_1^2 12a_1^2 13a_1^2 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^0$	3.46	3.54	3.26
	$^1A_2$	$11a_1^2 12a_1^2 13a_1^2 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^0$	5.17		
	$^1B_1$	$11a_1^2 12a_1^2 13a_1^2 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^1$	5.05		
	$^1B_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^2$	5.69		
	$^3A_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^1 7b_2^2 8b_2^0\ 2a_2^1$	4.73		
	$^3A_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	3.73	3.80	3.81
	$^3B_1$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^0$	3.85		3.81
	$^3B_2$	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^1\ 2a_2^0$	5.57		
	1^3B_2	$11a_1^2 12a_1^2 13a_1^0 14a_1^0\ 4b_1^2 5b_1^1\ 6b_2^2 7b_2^2 8b_2^0\ 2a_2^1$	5.30		
	2^3B_2	$11a_1^2 12a_1^2 13a_1^1 14a_1^0\ 4b_1^2 5b_1^0\ 6b_2^2 7b_2^2 8b_2^1\ 2a_2^0$	5.75		

[a] In the leading configuration, the $10a_1$, $3b_1$, $5b_2$, and $1a_1$ orbitals are always doubly occupied; the $15a_1$ orbital is always unoccupied.

detachment out of the $6b_2$, $4b_1$, and $12a_1$ orbitals of $^4B_2$, respectively. However, the corresponding VDEs for these ionizations are 4.82, 5.47, and 5.89 eV, and therefore they are much too high to explain the low-energy part of the photoelectron spectrum. Clearly, the X band cannot be explained by a one-electron ionization process from the $^4B_2$ state of FeO_3^-, and we need to explore other states of the anion.

Obviously, the next states of FeO_3^- that can be at the origin of its photoelectron spectroscopy are the nearly degenerate $^2A_2$ and $^2B_1$ components of the D_{3h} $^2E''$ state. Both the C_{2v} states have a slightly different equilibrium geometry, which explains the 0.12 eV vertical energy difference for $^2B_1$ in Table 5, which contains the CASPT2 energies from single point calculations on the geometry of the lower $^2A_2$ component. This contrasts with Table 4, which was obtained from the single point D_{3h} calculation using the Jahn−Teller inactive $^4B_2$ geometry. Starting from the $^2A_2$ and $^2B_1$ states, we can obtain the ground state $^1A_1$ of FeO_3 through a one-electron removal from the $2a_2$ and $5b_1$ orbitals, respectively. As advocated in a previous paragraph, the small energy barrier connecting the $^2A_2$ and $^2B_1$ states and the resulting vibronic coupling will give rise to a single band in the experimental spectrum, which we could describe as the D_{3h} transition $^2E'' \rightarrow {}^1A_1'$. Solely for computational convenience, we only included VDEs from the $^2A_2$ lowest energy component, which has certainly no bearing whatsoever on the proposed conclusions. The corresponding ADE and VDE values for this process are 3.37 and 3.46 eV, and we believe this transition is at the origin of the X band at 3.26 eV of the experimental spectrum. Moreover, if we remove one electron from the $13a_1$ orbital of $^2A_2$, we arrive at the $^3A_2$ state, for which the ADE and VDE values are 3.71 and 3.73 eV. These energies correspond very well with the A band at 3.81 eV. Judging from Table 5, we strongly believe that we could have drawn exactly the same conclusion from VDEs obtained for ionizations from the $^2B_1$ component. On the basis of all of the above arguments, we are inclined to make the following proposition about the photoelectron spectrum of FeO_3^-. The two low-lying bands should be assigned as originating from the $^2E''$ ($^2A_2$ and $^2B_1$) state of FeO_3^-, although our CASPT2 places this state at a higher energy of about 0.27 eV than $^4A_1'$. In order to get better values for the ADEs and VDEs, we performed single point CASPT2 calculations with larger ANO-RCC [8s,7p,6d,4f,2 g,1 h] and [7s,6p,4d,3f,1 g] basis sets for iron and oxygen, respectively. At this computational level, the VDE for the X feature is 3.54 eV, while for the A feature, we find 3.80 eV. $^4A_1'$ remains the ground state at this level, and the energy gap of 0.25 eV with $^2A_2$ ($^2E''$) is hardly affected.

Further evidence for the above proposed assignment can be obtained from the observed peak intensities in the vibrational progressions. As mentioned, the X feature is a broad progression of four peaks of relatively low intensity, which implies that there is a relatively larger difference between the geometries of the anionic FeO_3^- state and the final FeO_3 state. Otherwise, the A feature has a high-intensity sharp peak and a much lower second one, indicative of a smaller geometric difference between the geometries of two states that are responsible for this band. Also, the distinct vibrational progressions in the experimental spectrum strongly suggest that just one vibration mode lies at their origin. The rather large values of the associated vibrational frequencies of 850 cm^{-1} suggest a Fe−O stretching mode as observed for the diatomic FeO.[9−12,14,15,25] On this premise and using the potential energy curves of Figure 4, we can perform a harmonic vibration analysis for the symmetric stretching mode of the various low-lying states incorporated in this figure. The calculated Franck−Condon factors are depicted in Figure 7. For the lowest electron detachment, a vibrational frequency of 927 cm^{-1} was obtained for the $^1A_1'$ ground state of the neutral system, which corresponds well with the DFT value of 920 cm^{-1} from Table 1. Compared to the experimental 850 cm^{-1} from the photoelectron spectrum, it lies just outside the proposed error margin of 50 cm^{-1}. Further on, Figure 7a shows three Franck−Condon factors of more than 0.1, which agrees with the observed vibrational progression of the X band. Taking into account that our result is just a first estimate for the frequency, we interpret these results as a confirmation of the proposed assignment for this band. An even better correspondence was found for the A band. The calculated frequency of 881 cm^{-1} for the $^3A_2$ state is in

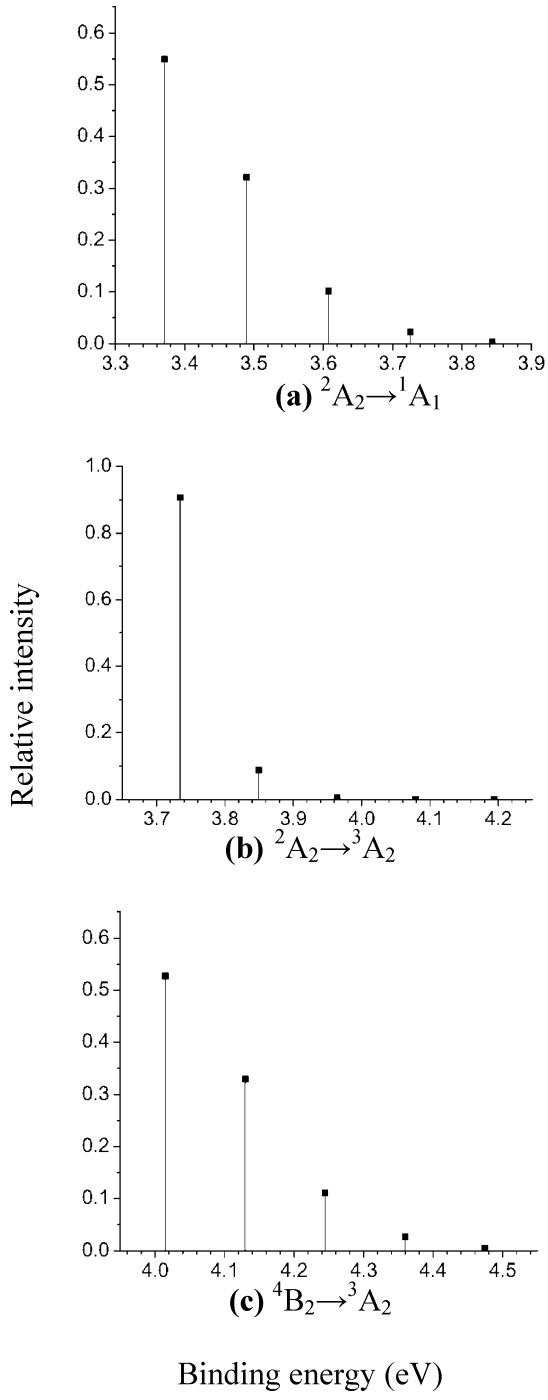


Figure 7. Simulated vibrational progression on the basis of harmonic Franck−Condon factors for the symmetric stretch of the three lowest energy electron detatchment processes. Peak positions derived from CASPT2 energies, BP/QZVP zero-point energies, and CASPT2 symmetric stretch frequencies.

complete agreement with the experimental value of 850 (50) cm^{-1} but somewhat at variance with the DFT result of 995 cm^{-1}. According to Figure 7b, the Franck−Condon factors for this band indicate just two observable peaks, in which the low energy one is about 10 times more intensive, a good match with the experimental spectrum of Figure 6. By studying the shapes of low-lying bands as above, we conclude that only the $^2E''$ lies at the origin of the photoelectron spectrum of FeO_3^-, and that detachments to the $^1A_1'$ and $^3E''$ states of the neutral complex are responsible for the X and A bands, respectively.

Conclusion

For the first time, the electronic structures of FeO_3 stoichiometry have been investigated at a multireference level of theory. We have found several stable geometries for both the singly charged anionic and neutral structures with different spin multiplicities. Regardless of the charge of these complexes, the η^2-O_2FeO conformations with O−O bonding are much higher in energy than the corresponding lowest iron triioxide conformations without any direct O−O bonding. Due to small Jahn−Teller effects, all of the equilibrium geometries of the latter conformation and the resulting photoelectron spectrum can effectively be described by using the planar D_{3h} symmetry. The computed CASSCF molecular orbitals as well as a qualitative interpretation of the relative CASPT2 energies unambiguously point to a formal oxidation state of +6 and +5 of iron in the neutral and anionic species, respectively. These oxidation states for both species imply that the ionization processes that underlie the bands observed in the experimental photoelectron spectrum correspond formally to a detachment of a metal electron. However, a population analysis of the ab initio wave functions shows that, most likely, the observed ionization processes involve the removal of electron density from the oxygen atoms. Quite remarkable, a general analysis in terms of a simply ligand field description of the splitting pattern of the valence iron 3d orbitals is effective to describe the electronic structure of the clusters and their studied spectroscopy. The ground state of the neutral cluster is the closed shell $^1A_1'$, but the lowest state of the anion can be either the strong-field ground state or low-spin $^2E''$ ($^2A_2$, $^2B_1$) or the weak-field ground state or high-spin $^4A_1'$. A 0.25 eV energy gap between them is judged to be too small to conclude unequivocally which is the true ground state of the anion, although the latter state should be seen as a serious candidate. From our CASPT2 calculations, we are inclined to assign the low-lying bands of the photoelectron spectrum of FeO_3^- as electron detachment processes from the low-spin $^2E''$ state. Our best VDEs (ADEs) values are 3.54 (3.44) eV and 3.80 (3.77) eV, which compare convincingly well with the start position of the X band at 3.26 eV and of the A band at 3.81 eV. The closed shell ground state $^1A_1'$ and the first excited state $^3E''$ of the neutral cluster lie at the origin of these two bands. Analyzing the vibrational progression of these transitions further substantiates our proposed assignment.

Acknowledgment. This study has been supported by grants from the Flemish Science Foundation (FWO) and from the Concerted Research Action of the Flemish Government (GOA).

References

(1) Gong, Y.; Zhou, M.; Andrews, L. *Chem. Rev.* **2009**, *109*, 6765.

(2) Jena, P.; Castleman, A. W., Jr. *Proc. Natl. Sci. U.S.A.* **2006**, *103*, 10560.

(3) Mebel, A. M.; Hwang, D. Y. *J. Phys. Chem. A* **2001**, *105*, 7460.

(4) Xue, W.; Wang, Z. C.; He, S.-G.; Xie, Y.; Bernstein, E. R. *J. Am. Chem. Soc.* **2008**, *30*, 15879.

(5) El-Sheikh, S. M.; Harraz, F. A.; Abdel-Halim, K. S. *J. Alloy Compd.* **2009**, *487*, 716.

(6) Fellah, M. F.; Onal, I.; van Santen, R. A. *J. Phys. Chem. C.* **2010**, *114*, 12580.

(7) Gutsev, G. L.; Bauschlicher, C. W., Jr. *Chem. Phys. Lett.* **2003**, *380*, 435 .

(8) Reddy, B. V.; Khanna, S. N. *Phys. Rev. Lett.* **2004**, *93*, 68301.

(9) Fan, J.; Wang, L.-S. *J. Chem. Phys.* **1995**, *102*, 8714.

(10) Wu, H.; Desai, S. R.; Wang, L.-S. *J. Am. Chem. Soc.* **1996**, *118*, 5296.

(11) Wu, H.; Desai, S. R.; Wang, L.-S. *J. Am. Chem. Soc.* **1996**, *118*, 7434.

(12) Wang, L.-S.; Wu, H.; Desai, S. R. *Phys. Rev. Lett.* **1996**, *76*, 4853.

(13) Chertihin, G. V.; Saffel, W.; Yustein, J. T.; Andrews, L.; Neurock, M.; Ricca, A.; Bauschlicher, C. W., Jr. *J. Phys. Chem.* **1996**, *100*, 5261.

(14) Gutsev, G. L.; Khanna, S. N.; Rao, B. K.; Jena, P. *J. Phys. Chem. A* **1999**, *103*, 5812.

(15) Gong, Y.; Zhou, M. *J. Phys. Chem. A* **2008**, *112*, 10838.

(16) Uzunova, E. L.; Milosch, H.; Nikolov, G. S. *J. Chem. Phys.* **2008**, *128*, 94307.

(17) Clima, S.; Hendrickx, M. F. A. *Chem. Phys. Lett.* **2007**, *436*, 341.

(18) Hübner, O.; Volker, T.; Berning, A.; Sauer, *J. Chem. Phys. Lett.* **1998**, *294*, 37.

(19) Clima, S.; Hendrickx, M. F. A. *J. Phys. Chem. A* **2007**, *111*, 10992.

(20) Gutsev, G. L.; Rao, B. K.; Jena, P. *J. Phys. Chem. A* **2000**, *104*, 5374.

(21) Grein, F. I. *Int. J. Quantum Chem.* **2009**, *109*, 549.

(22) Garcia-Sosa, A. T.; Castro, M. *Int. J. Quantum Chem.* **2000**, *80*, 307.

(23) Schröder, D.; Kretzschaar, I.; Schwarz, H.; Rue, C.; Armentrout, P. B. *Inorg. Chem.* **1999**, *38*, 3474.

(24) Hübner, O.; Sauer, J. *J. Chem. Phys.* **2002**, *116*, 617.

(25) Hendrickx, M. F. A.; Anam, K. R. *J. Phys. Chem. A* **2009**, *113*, 8746.

(26) Cardoen, W.; Gdanitz, R. J. *Chem. Phys. Lett.* **2002**, *364*, 39.

(27) Li, Z. H.; Gong, Y.; Fan, K.; Zhou, M. *J. Phys. Chem. A* **2008**, *112*, 13641.

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

(29) Ahlrichs, R.; Furche, F.; Grimme, S. *Chem. Phys. Lett.* **2000**, *325*, 317.

(30) *TURBOMOLE*, V6.1 2009; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 1989, TURBOMOLE GmbH since 2007. Available from http://www.turbomole.com (accessed Dec 9, 2010).

(31) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neograd y, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.

(32) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A*. **2005**, *108*, 2851.

(33) Douglas, N.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89.

(34) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742.

(35) Gutsev, G. L.; Khanna, S. N.; Rao, B. K.; Jena, P. *Phys. Rev. A* **1999**, *59*, 3681.

(36) Gutsev, G. L.; Weatherford, C. A.; Pradhan, K.; Jena, P. *J. Phys. Chem. A* **2010**, *114*, 9014.

CT1005246

JCTC Journal of Chemical Theory and Computation

Accurate Quantum Chemistry in Single Precision Arithmetic: Correlation Energy

Victor P. Vysotskiy* and Lorenz S. Cederbaum

Theoretical Chemistry, Institute of Physical Chemistry at Heidelberg University, Im Neuenheimer Feld 229, 69120 Heidelberg, Germany

Received September 17, 2010

Abstract: In the present work, we show the feasibility of using single precision in quantum chemistry, especially regarding the computation of electron correlation energy. On the example of the MP2 method, we clearly demonstrate that single precision arithmetic is sufficient for evaluating the molecular two-electron integrals by the use of the Cholesky decomposition method. The evaluation of integrals with single precision arithmetic introduces a negligible error into the computed MP2 correlation energy. In particular, the corresponding error in the MP2 correlation energy amounts to only $10^{-7}E_h$ for the 113-atom taxol molecule in double-valence basis set (1099 basis functions). The practical relevance of our result is that 50% performance gain and 50% reduction in memory demands can be achieved by only minor changes in the existing codes. Our finding opens intriguing perspectives for doing accurate correlated quantum chemistry on specialized floating-point mathematical coprocessors.

1. Introduction

The use of double precision is the most common convention in quantum chemistry.[1] This programming rule is biased by the believed paradigm that higher precision automatically yields more accurate results. The interrelation between the precision and the accuracy of the final results is not straightforward, however.[2,3] Indeed, the accuracy of the final results depends not only on the precision used but also on many factors like, e.g., the algorithm utilized, intermediate data generated, compilers and math libraries employed, the hardware architecture, and so forth.

In computational quantum chemistry, the use of double precision is essential at the stages of the implementation and validation of a new theoretical method or computational scheme. Once the method is approved and the code is verified, usually by examples of small and medium-sized systems, further modifications are necessary to enable large scale (many atoms and many basis functions) *ab initio* calculations. At the present time, there exists a wide choice of powerfull methods which can help here: three-index factorization of two-electron repulsion integrals (ERIs), local schemes, the fragmented molecular orbital (FMO) method, and many others.[4−7] These accelerating methods have to provide an optimal balance between the accuracy of the computed quantities and the computational effort required.

Since the formal scaling of the number of ERIs is quartic with respect to the number of basis functions, the evaluation of ERIs is the major computational obstacle in all advanced quantum chemistry calculations. A three-index factorization of the ERIs is known to be a very efficient technique for reducing the computational prefactor and to speed up ab initio calculation. Such a factorization is the cornerstone of the Density Fitting or Resolution of Identity (DF/RI) and Cholesky Decomposition (CD) methods.[8−14] In the framework of this factorization, an ERI is approximated by the inner product of two intermediate vectors (ERI is expressed in terms of three-indexed intermediates). According to published results, the absolute error in an ERI caused by the approximation lies within the range of 10^{-2} to 10^{-16} E_h, depending on the method used.[13,15−17]

By applying the DF/RI and CD methods, one can accelerate ab initio calculations up to a few hundred times and thereby simulate large quantum systems in a reasonable time.[18,19] This acceleration comes, of course, at the cost of accuracy: using approximated ERIs leads to deterioration of the numerical accuracy of the final results (energies, proper-

* To whom correspondence should be addressed. E-mail: victor.vysotskiy@pci.uni-heidelberg.de.

10.1021/ct100533u
Published on Web 12/23/2010

ties), even though calculations are carried out using double precision. In other words, all of the above-mentioned approximation methods introduce a systematic error to the computed quantities. According to the present standard, this *approximation* error must not exceed the so-called *chemical accuracy*, which is defined to be 1 kcal/mol or, equivalently, 1.593×10^{-3} E_h.[20,21]

From the numerical point of view, the factorization might be effectively interpreted as rounding an exact ERI from double to some intermediate precision. The loss of numerical accuracy resulting from integral approximation opens up the possibility for the use of single precision during the internal intermediate calculations. One might speculate that an energy error caused by working with three-index intermediates in single precision mode (storage and computation) is comparable to the approximation error or is even smaller. The computational benefits of using single precision are enormous. First of all, it automatically halves the memory demands and doubles memory and network bandwidths. Second, single precision arithmetic (32-bit arithmetic) is at least 2× times faster on conventional processors (x86, x86−64, Intel 64, IA-64, IBM Power) and 10× times faster (!) on specially designed mathematical coprocessors (Nvidia's and AMD/ATI's General Purpose Graphics Processing Unit, IBM's Cell BE) than double precision arithmetic (64-bit arithmetic).[22−24]

It is thus not surprising that in recent years the use of single precision in quantum chemistry has attracted considerable attention, especially regarding the evaluation of ERIs. Two computational schemes have been already implemented and assessed: computation of ERIs in an atomic orbital basis for direct HF and DFT calculations[25−30] and evaluation of ERIs in a molecular orbital basis with the DF/RI method for calculating the so-called RI-MP2 correlation energy.[31−33] At present, the prevailing opinion concerning the evaluation of ERIs in single precision arithmetic is that "single precision is generally insufficient to achieve 'chemical accuracy' of 1 kcal/mol in calculations on anything but the smallest and simplest systems, since the errors quickly accumulate for large molecules."[33]

Indeed, the accumulated error in the total Hartree−Fock (HF) and correlation energies grows rapidly with system size and becomes unacceptable ($\geq 1.593 \times 10^{-3}$ E_h) for moderate-sized molecules ($\sim 10^2$ atoms and $\sim 10^3$ basis functions).[26,31] In order to overcome this problem, a mixed precision computational model was developed. This model utilizes both the host CPU (for evaluating large ERIs with 64-bit arithmetic) and an attached GPGPU (for the evaluation of the remaining of ERIs with 32-bit arithmetic).[28,32] By using this CPU−GPU heterogeneous model, the required accuracy of 1 kcal/mol has been achieved.[27,33] In particular, Aspuru-Guzik and co-workers, by using this mixed-precision computational model, have reduced the absolute error in the RI-MP2 correlation energy from 9.980×10^{-3} to 7.986×10^{-4} E_h for the 113-atom taxol molecule in a double-ζ valence basis set (1123 basis functions).[32,33]

The aim of the present study is to demonstrate the particular feasibility and practicability of using single precision in conjunction with three-indexed intermediates generated via the CD method. In contrast to the DF/RI method, the potential of the CD method for generating ERIs within single precision arithmetic has never been investigated before. Our computational strategy is to focus here on the correlated level of theory.

2. Theoretical Background

2.1. The Three-Index Factorization of Two-Electron Repulsion Integrals. Generally, an ERI in the framework of a three-index factorization can be represented as the inner product of two intermediate vectors:

$$(\mu\nu|\lambda\sigma) \approx \widetilde{(\mu\nu|\lambda\sigma)} = \mathbf{L}_{\mu\nu} \cdot \mathbf{L}_{\lambda\sigma} = \sum_{K=1}^{M} L_{\mu\nu}^{K} L_{\lambda\sigma}^{K} \qquad (1)$$

where μ, ν, λ, and σ label atomic orbitals; $(\mu\nu|\lambda\sigma)$ and $\widetilde{(\mu\nu|\lambda\sigma)}$ are the exact ERI and its approximation in Mulliken notation, respectively; $L_{\lambda\sigma}^{K}$ are three-indexed intermediates. In the particular case of the CD method, $\mathbf{L}_{\mu\nu}$ is called the *Cholesky vector* in the AO basis and M is the number of Cholesky vectors.

The main advantage of the CD method is that the accuracy of the approximation (eq 1) can be rigorously controlled. By construction, the accuracy control is accomplished by varying only one parameter, the so-called CD threshold δ:

$$\Delta = |(\mu\nu|\lambda\sigma) - \widetilde{(\mu\nu|\lambda\sigma)}| \leq \delta \qquad (2)$$

where Δ is the approximation error of an ERI. Depending on the decomposition scheme, the strict error bound (eq 2) holds for all ERIs or only for certain types. In the case of the full-CD scheme, the introduced error can be made as small as needed for all ERIs:

$$\lim_{\delta \to \infty} \Delta = \varepsilon$$

or equivalently

$$\lim_{\delta \to \infty} \widetilde{(\mu\nu|\lambda\sigma)} = (\mu\nu|\lambda\sigma) \qquad (3)$$

where ε is the machine epsilon (2.220×10^{-16}). In finite precision arithmetic, this limit is reached when $\delta \leq 10^{-10}$. In the case of the recently developed *atomic CD* (aCD) or its compact form (acCD), this inequality (eq 2) is valid only for the one-center and two-center "Coulomb" ERIs, but three- and four-center integrals as well as "exchange" two-center integrals may be subject to large errors.[15,16,34] In other words, within the framework of the aCD/acCD schemes, the accuracy of the approximation (eq 1) for a part of ERIs cannot be improved beyond a certain limit no matter what CD threshold is used.

In practice, a CD threshold in the range of 10^{-4} to 10^{-6} is being used in most applications, and the corresponding number of the Cholesky vectors (M) is 5−7 times larger than the number of the basis functions (N). These CD thresholds guarantee the chemical accuracy of the final results and accelerate the calculation up to a few hundred times. By taking into account eq 2 and these CD thresholds, one might claim that the actual precision of the approximated ERIs used is numerically close to single

(8 significant decimal digits) rather than double (16 significant decimal digits) precision.

Another concern related to using the CD method is the final error in the computed energies. As rule of thumb, the absolute error of computed total energies and other properties caused by using the CD method is proportional to δ and becomes virtually equal to zero for $\delta \leq 10^{-10}$. Relative energies like, e.g., electron propagator poles (ionization potentials and electron affinities) or excitation energies are very robust with respect to the CD threshold and converge rapidly to the numerically exact ones ($\delta \geq 10^{-5}$ is more than enough to achieve millielectronvolt accuracy).[35,36]

Before we leave this section, let us briefly discuss some technical aspects of the factorization. For correlated methods, an important feature of this factorization is that it holds also in the case of the molecular orbital (MO) representation. If **C** is the MO expansion coefficients matrix, then an ERI in MO representation can be calculated by exploiting the same factorization (eq 1):

$$(pq|rs) = \sum_{K=1}^{M} L_{pq}^{K} L_{rs}^{K} \tag{4}$$

where p, q, r, and s denote MO indices and a MO transformed Cholesky vector L_{pq}^{K}:

$$L_{pq}^{K} = \sum_{\mu} \sum_{\nu} C_{\mu p} L_{\mu\nu}^{K} C_{\nu q} \tag{5}$$

reduces the scaling of the atomic orbital (AO) to molecular orbital (MO) transformation from $\mathcal{O}(N^5)$ to $\mathcal{O}(N^4)$. The CD factorization substantially reduces storage demands (by factor $\sim N^2/M$) and I/O overheads and thus converts the problem of determining electron correlation energies from a memory-bound one to a compute-bound one.[37,38] The time needed to complete a compute-bound task depends mostly on the performance of an execution unit (CPUs core, GPUs core) and can be significally reduced by using single precision arithmetic.

2.2. A Priori Error Estimation Caused by Using Single Precision Arithmetic. In MP2 theory, only the $(ov|ov)$ class of molecular ERIs is needed to compute the electron correlation energy $E^{(2)}$:

$$E^{(2)} = \sum_{\substack{i \\ j \geq i}}^{N_o} \sum_{\substack{a \\ b}}^{N_v} \frac{(2 - \delta_{ij})[2(ai|bj) - (aj|bi)](ai|bj)}{\varepsilon_i - \varepsilon_a + \varepsilon_j - \varepsilon_b} \tag{6}$$

where a and b denote virtual orbitals, and i and j denote occupied HF orbitals; N_o and N_v are the total number of ocuppied and virtual orbitals, respectively. Let us estimate the error introduced by using single precision arithmetic for generating an approximated ERI via formula eq 4.

The error introduced in the inner product due to single precision arithmetic can be estimated as follows:[39,40]

$$|fl(\widetilde{ai|bj}) - (\widetilde{ai|bj})| \leq \gamma_M \sqrt{(\widetilde{ai|ai})(\widetilde{bj|bj})} \tag{7}$$

where

$$\gamma_M = \frac{Mu}{1 - Mu} \tag{8}$$

$fl(\widetilde{ai|bj})$ means that an approximated ERI is computed with single precision arithmetic via eq 4, and $u = 2^{-24} \approx 5.960 \times 10^{-8}$ is named *unit roundoff*.[41] The prefactor γ_M depends hyperbolically on M. However, as long as $M \leq 10^6$, this dependency is essentially linear. Since the number of Cholesky vectors M typically grows linearly with system size, the prefactor γ_M should scale linearly with the system size. By looking at the square root in eq 7, one can easily recognize the so-called Schwarz upper bound to an ERI. Figure 1 shows the distribution of the upper bounds of the ERIs relevant to the MP2 correlation energy for the illustrative example of the water dimer $(H_2O)_2$ according to eq 7. We notice that this distribution is in the range from 10^{-5} to 10^{-8} and peaks at 10^{-6}.

According to our empirical experience, this is a typical distribution. A relevant reason for this well behaved distribution is that the Cholesky vectors are free from large components (see Figure S1 in the Supporting Information). The CD method is numerically well conditioned, and available implementations are very robust.[14,19,42] CD methods have other appealing features related to the structure of the Cholesky vectors which we do not discuss here, and we refer the interested reader to refs 16, 34, and 43 for more details.

A particularly relevant point is that the upper bounds of the ERIs shown in Figure 1 significally overestimate the true error induced by the use of single precision arithmetic. For this purpose, we also show in Figure 1 the distribution of the true errors of the ERIs, i.e., of $|fl(\widetilde{ai|bj}) - (\widetilde{ai|bj})|$. As clearly seen, the true errors range from 10^{-8} to 10^{-14}, and the distribution peaks at about 10^{-11}.

In summary, we expect that the evaluation of the ERIs in single precision arithmetic has only a slight impact on the MP2 correlation energies. To be more precise, we claim that the error caused by single-precision arithmetic is expected to be comparable to the error introduced by the CD method in common practical computations.

3. Computational Details

For test calculations, we used a set of water clusters $(H_2O)_n$ ($n = 2, ..., 20$)[44,45] and the taxol molecule ($C_{47}H_{51}NO_{14}$).[46] In the present study, we employed Roos's ANO-L-VXZP (X = D, T) basis sets.[47,48]

The MP2 method in single precision was implemented in the development version of the P-RICDΣ program.[35] As input data, P-RICDΣ uses the integral tables in the AO representation ($\mathbf{L}_{\mu\nu}$) and the SCF MO LCAO coefficients which are generated with the MOLCAS ab initio package.[19,49]

Within P-RICDΣ, the computation of the MP2 correlation energy proceeds in two steps: a stepwise parallel transformation of integral tables from the AO to MO representation and subsequent calculation of the $E^{(2)}$ energy correction via eq 6. The transformed Cholesky vectors in the MO basis ($\mathbf{L}_{ai}$) are stored in single precision (each number occupies 32 bits rather than 64 bits). The approximated ERIs are

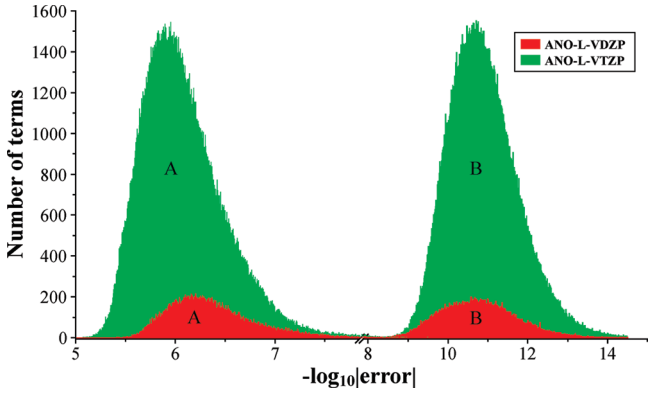


Figure 1. Distributions of the errors of the $(ai|bj)$ ERIs for the water dimer using ANO-L-VDZP and ANO-L-VTZP atomic basis sets: (A) The distribution of the upper bounds according to inequality 7, i.e., $\gamma_M\sqrt{(\widetilde{ai|ai})(\widetilde{bj|bj})}$. (B) The distribution of the true error, i.e., $|fl(\widetilde{ai|bj}) - (\widetilde{ai|bj})|$. Note the logarithmic scale used.

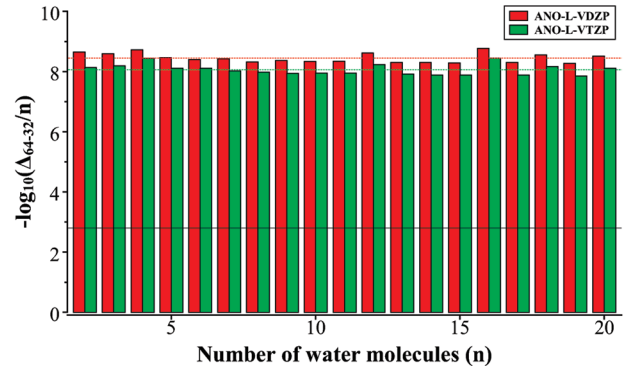


Figure 2. The normalized absolute error (Δ_{64-32}/n) in the all-electron MP2 energy caused by single precision for a series of water clusters $(H_2O)_n$ ($n = 2-20$) employing the ANO-L-VDZP and ANO-L-VTZP basis sets. The dash horizontal lines display the average normalized error. Those are 3.55×10^{-9} (red color) and 8.65×10^{-9} E_h (green color) in the VDZP and VTZP basis sets, respectively. The solid horizontal line (black) indicates chemical accuracy (1.593×10^{-3} E_h). Note the logarithmic scale used.

computed in single precision arithmetic by calling an appropriate *sdot* BLAS function (single precision inner product). The above-described utilization of single precision (type casting and using the *sdot* function) automatically reduces the memory demands and execution time by a factor of 2. In order to accumulate the $E^{(2)}$ energy corrections, double precision was used because this yields a substantial gain in the final accuracy. Summation in double precision which scales as $\mathcal{O}(N_o^2N_v^2)$ does not lead to any performance degradation because it only constitutes less than 1% ($\propto 1/M$) of the total number of floating point operations required. The most computationally demanding step of the entire algorithm is by far the generation of the $(\widetilde{ai|bj})$ ERIs, which scales as $\mathcal{O}(N_o^2N_v^2M)$.[31,50]

For comparison, standard MP2 energy calculations (in double precision) were carried out within the MOLCAS 7 program using a CD-based implementation.[51] It should be particularly emphasized that the single precision MP2 calculations were done over exactly the same data and in the same runtime environment that were used for the double precision ones.

All programs used in this work were compiled within Intel Cluster Toolkit 4.0. All calculations were done on Intel Xeon E5440 (2.83 GHz) processors.

In all calculations, a CD threshold of 10^{-10} was used unless otherwise specified.

4. Results and Discussion

4.1. Water Clusters. The key results of our work are displayed in Figure 2 (see also Figure S2 and Tables S3 and S4 in the Supporting Information). The figure shows the absolute error in the $E^{(2)}$ energy correction (Δ_{64-32}) normalized to the size of the cluster (i.e., divided by the number of water molecules) caused by the evaluation of ERIs in single precision. The error is determined as the absolute difference between the MP2 energy computed using the P-RICDΣ (single precision) and MOLCAS 7 (double precision) programs and the $\delta = 10^{-10}$ CD threshold. As predicted in section 2, single precision arithmetic introduces a negligible

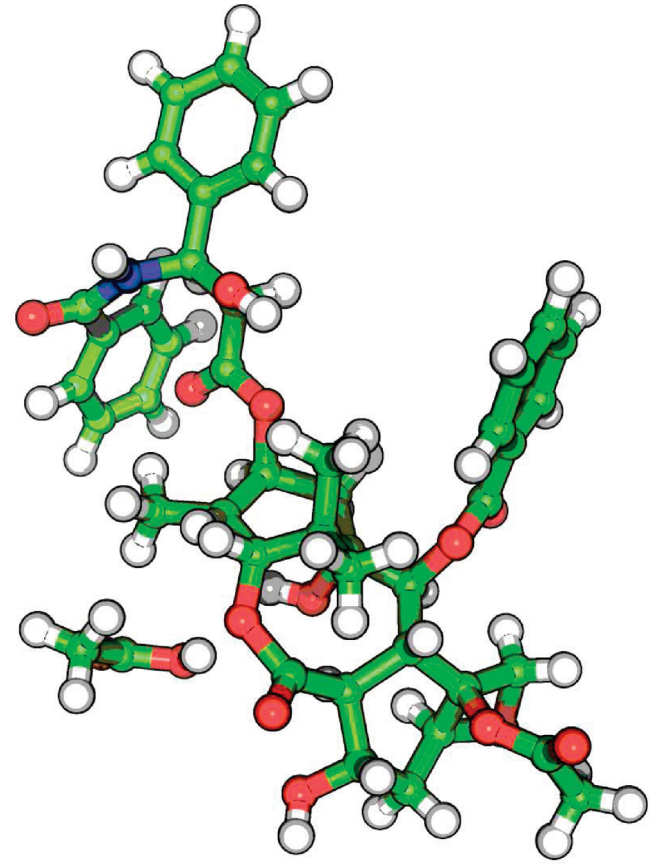


Figure 3. The taxol molecule ($C_{47}H_{51}NO_{14}$). The total number of basis functions in the ANO-L-VDZP basis set is 1099.

error into the computed MP2 correlation energy. The actual error encountered in the correlation energies amounts to 100 nHartree (1 nHartree $= 10^{-9}$ E_h) only. In particular, the maximal (mean) absolute errors in the $E^{(2)}$ energy correction are 100 (43.7) and 269 (107) nHartree in the ANO-L-VDZP and ANO-L-VTZP basis sets, respectively.

It is clearly seen from Figure 2 that the normalized error only slightly depends on the cluster size and mostly depends

Table 1. The $E^{(2)}$ Correlation Energies for the Taxol Molecule Computed in Double ($E_{64}^{(2)}$) and Single Precision ($E_{32}^{(2)}$) by Employing Different CD Schemes and the ANO-L-VDZP Valence Basis Set[a]

method[b]	$E_{64}^{(2)}$	$E_{32}^{(2)}$	Δ_{CD}	Δ_{64-32}
CD-4	−9.399108640	−9.399108549	1.906×10^{-3}	9.100×10^{-8}
acCD-4*	−9.400377074	−9.400377031	6.373×10^{-4}	4.300×10^{-8}
acCD-4	−9.396719436	−9.396719405	4.295×10^{-3}	3.100×10^{-8}
CD-6	−9.400953079	−9.400952911	6.128×10^{-5}	1.680×10^{-7}
acCD-6*	−9.400673749	−9.400673704	3.406×10^{-4}	4.500×10^{-8}
acCD-6	−9.397581736	−9.397581697	3.433×10^{-3}	3.900×10^{-8}
CD-9	−9.401014359	−9.401013997	0.000	3.620×10^{-7}

[a] The CD-9 scheme is highly accurate and provides the reference value. The basis set contains 1099 basis functions; the number of the doubly occupied orbitals is 223. Shown also are the absolute errors of the correlation energies caused by using single precision arithmetic (Δ_{64-32}) and the errors of the various approximate CD schemes (Δ_{CD}). All energies in au. [b] CD-*n* refers to the full-CD decomposition threshold $\delta = 10^{-n}$. acCD-*n** means "*atomic compact CD*", and acCD-*n* auxiliary basis sets have been formed from the original acCD-*n** ones by removing the highest angular momentum functions. See refs 15, 16 for details.

on the atomic basis set used. By going from the ANO-L-VDZP to the ANO-L-VTZP basis set, the averaged normalized error changes from 3.6 to 8.7 nHartree, i.e., becomes only 2.42 times larger. At the same time, the number of floating point operations needed to compute $E^{(2)}$ increases by a factor of ∼18.62 (the generation of the $(\widetilde{ai}|\widetilde{bj})$ ERIs). This difference by about one order of magnitude is due to the cancellation of errors when summing up the contributions from the individual integrals in eq 6. By taking into account the information that the corresponding number of Cholesky vectors increased by a factor of 2.43, in average, over the set of water clusters (see Table S5 in the Supporting Information), we can claim that the Δ_{64-32} error varies linearly with the number of Cholesky vectors, i.e., with the system size. This observation is in complete agreement with the error model considered in the section 2.2 (see eqs 7 and 8).

Let us make a rough estimate of the critical size which a system must have in order to cross the limit of chemical accuracy. By taking the value 10 nHartree (10^{-8} E_h) as averaged error per water molecule, the critical size is estimated to be 160 000 water molecules or, equally, 500 000 atoms. This critical size is currently much beyond reach for correlated quantum chemistry.

4.2. Taxol Molecule. In order to demonstrate that the numerical results reported above are general and are not biased to the water clusters set only, let us consider another example, namely, the taxol molecule (see Figure 3). Table 1 reports the MP2 correlation energies computed in double and single precision by employing various CD schemes. As in the case of the water clusters, the single precision errors (Δ_{64-32}) are negligibly small: the error lies in the range from 31 to 362 nHartree. It is clearly seen from the Table 1 that the error caused by the use of single precision (Δ_{64-32}) is a few orders of magnitude smaller than the corresponding approximation error Δ_{CD} of the CD scheme.

5. Future Prospect and Perspectives

The high-end floating-point mathematical coprocessors available on the market offer teraflop (10^{12} floating point operations per second) single-precision performance. For example, the performance of Nvidia's Tesla S2050 GPGPU and IBM's PowerXCell 8i based solutions are 4.1 and 6.4 teraflops, respectively.[52,53] Such performances are roughly equivalent to the total performance of 200 Intel Xeon 54xx (Harpertown) cores. But the current price of either Nvidia's or IBM's solution is only a 1/10th that of the corresponding CPUs. With respect to the results we have obtained, we consider coprocessors as very promising computational platforms for performing accurate large-scale correlated calculations. With particular emphasis on electron propagator calculations, which are our primary goal, we plan to extend the capabilities of our P-RICDΣ program accordingly (transfer Cholesky vectors to coprocessors and generate the needed molecular integrals via a BLAS library provided by vendor).

6. Conclusions

In the present work, we clearly demonstrate by the illustrative example of MP2 theory that single precision is sufficient for post Hartree−Fock methods relying on the Cholesky decomposition of the two-electron integrals. The key advantage of the proposed scheme is that 50% performance gain and 50% reduction in memory demands can be achieved by only minor changes in the existing codes. Our results open intriguing perspectives for future developments and trends in the computational quantum chemistry.

Acknowledgment. Financial support by the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged. We would like to thank bwGrid[54] for providing computational resources.

Supporting Information Available: Graphical representation of the distrubution of the components of Cholesky vectors in molecular orbital basis sets. The graphical representation of the absolute error (Δ_{64-32}) in the all-electron MP2 energy caused by single precision, tables containing total Hartree−Fock energies, MP2 correlation energies (computed with double and single precision), number of cholesky vectors for a set of $(H_2O)_n$ ($n = 2, ..., 20$) water clusters employing the CD threshold of $\delta = 10^{-10}$ and the ANO-L-VDZP and ANO-L-VTZP atomic basis sets. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) In the source codes of the most widely used ab initio packages such as Molcas, Molpro, and ACES-III, one can easily detect

that each function or subroutine explicitly declares all variables in double precision.

(2) Boisvert, R. F.; Cools, F.; Einarsson, B. Precision, Accuracy and Reliability. In *Accuracy and Reliability in Scientific Computing (Software, Environments, Tools)*; Einarsson, B., Ed.; SIAM Press: Philadelphia, PA, 2005; Vol. SE-18, p 21.

(3) Loh, E.; Walster, G. W. *Reliab. Comput.* **2002**, *8*, 245–248.

(4) Martinez, T. J.; Carter, E. A. Pseudospectral methods applied to the electron correlation problem. In *Modern Electronic Structure Theory. Part II*, 2nd ed.; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Vol. 2, pp 1132−1165.

(5) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151–154.

(6) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213–236.

(7) Fedorov, D. G.; Ishimura, K.; Ishida, T.; Kitaura, K.; Pulay, P.; Nagase, S. *J. Comput. Chem.* **2007**, *28*, 1476–1484.

(8) Harris, F. E.; Rein, R. *Theor. Chim. Acta* **1966**, *6*, 73–82.

(9) Billingsley, F. P.; Bloor, J. E. *J. Chem. Phys.* **1971**, *55*, 5178–5190.

(10) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(11) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(12) Linderberg, J. *Int. J. Quant. Symp.* **1977**, *S11*, 353–357.

(13) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.

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

(15) Aquilante, F.; Lindh, R.; Pedersen, T. B. *J. Chem. Phys.* **2007**, *127*, 114107.

(16) Aquilante, F.; Gagliardi, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2009**, *130*, 154107. We note that the Supplementary Material corresponding to this article contains important information.

(17) Aquilante, F.; Pedersen, T. B.; Lindh, R.; Roos, B. O.; Sánchez de Merás, A.; Koch, H. *J. Chem. Phys.* **2008**, *129*, 024113.

(18) Hättig, C. Beyond Hartree-Fock: MP2 and Coupled Cluster Methods for Large Systems. In *Computational Nanoscience: Do It Yourself!*; Grotendorst, J., Blügel, S., Marx, D., Eds.; John von Neumann Institute for Computing: Jülich, Germany, 2006; Vol. 31, pp 245−278.

(19) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224–247.

(20) Pople, J. A. *Rev. Mod. Phys.* **1999**, *71*, 1267–1274.

(21) Hoffmann, R.; Schleyer, P. V. R.; Schaefer, H. F., III. *Angew. Chem.* **2008**, *47*, 7164–7167.

(22) Tesla C2050/C2070 GPU Computing Processor. http://www.nvidia.com/object/product_tesla_C2050_C2070_us.html (accessed Sep 17, 2010).

(23) AMD FireStream Technology, see for example: AMD FireStream 9270 GPU Compute Accelerator. http://www.amd.com/us/products/workstation/firestream/firestream-9270/ Pages/firestream-9270.aspx (accessed Sep 17, 2010).

(24) The Cell project at IBM Research. http://www.research.ibm./cell/cell_chip.html (accessed Sep 17, 2010).

(25) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.

(26) Ufimtsev, I. S.; Martinez, T. J. *Comput. Sci. Eng.* **2008**, *10*, 26–34.

(27) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.

(28) Yasuda, K. *J. Comput. Chem.* **2007**, *29*, 334–342.

(29) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230–1236.

(30) Asadchev, A.; Allada, V.; Felder, J.; Bode, B. M.; Gordon, M. S.; Windus, T. L. *J. Chem. Theory Comput.* **2010**, *6*, 696–704.

(31) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049–2057.

(32) Olivares-Amaya, R.; Watson, M. A.; Edgar, R. G.; Vogt, L.; Shao, Y.; Aspuru-Guzik, A. *J. Chem. Theory Comput.* **2010**, *6*, 135–144.

(33) Watson, M.; Olivares-Amaya, R.; Edgar, R. G.; Aspuru-Guzik, A. *Comput. Sci. Eng.* **2010**, *12*, 40–51.

(34) Pedersen, T. B.; Aquilante, F.; Lindh, R. *Theor. Chem. Acc.* **2009**, *124*, 1–10.

(35) Vysotskiy, V. P.; Cederbaum, L. S. *J. Chem. Phys.* **2010**, *132*, 044110.

(36) Böstrom, J.; Delcey, M. G.; Aquilante, F.; Serrano-Andrés, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Theory Comput.* **2010**, *6*, 747–754.

(37) Memory bound refers to a situation in which the time to complete a given computational problem is decided primarily by the amount of available memory to hold data. In other words, the limiting factor of solving a given problem is the memory access speed.

(38) CPU bound (or compute bound) is when the time for a computer to complete a task is determined principally by the speed of the central processor.

(39) Higham, N. J. Basics. In *Accuracy and Stability of Numerical Algorithms*, 2nd ed.; SIAM Press: Philadelphia, PA, 2002; pp 62−65.

(40) Golub, G. H.; Van Loan, C. F. Matrix analysis. In *Matrix Computations*, 3rd ed.; The Johns Hopkins University Press: Baltimore, MD, 1996; pp 62−64.

(41) IEEE standard for binary floating-point arithmetic. ANSI/IEEE Standard, Std 754−1985, New York, 1985.

(42) Røeggen, I.; Johansen, T. *J. Chem. Phys.* **2008**, *128*, 194107.

(43) Jung, Y.; Sodt, A.; Gill, P. M. W.; Head-Gordon, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6692–6697.

(44) Maheshwary, S.; Patel, N.; Sathyamurthy, N.; Kulkarni, A. D.; Gadre, S. R. *J. Phys. Chem. A* **2001**, *105*, 10525–10357.

(45) The Cambridge Cluster Database, Ab initio Optimized $(H2O)_N$ Clusters. http://www-wales.ch.cam.ac.uk/wales/CCD/anantwatcl.html (accessed Sep 17, 2010).

(46) The cartesian coordinates are available on the Web: http://www.petachem.com/data/taxol.xyz (accessed Sep 17, 2010).

(47) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. *Theor. Chim. Acta.* **1990**, *77*, 291–306.

(48) Widmark, P.-O.; Persson, B. J.; Roos, B. O. *Theor. Chim. Acta.* **1991**, *79*, 419–432.

(49) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(50) Katouda, M.; Nagase, S. *Int. J. Quantum Chem.* **2009**, *109*, 2121–2130.

(51) Aquilante, F.; Pedersen, T. B. *Chem. Phys. Lett.* **2007**, *449*, 354–357.

(52) The Tesla S2050 1U Computing System. http://www.nvidia.com/object/product-tesla-S2050-us.html (accessed Sep 17, 2010).

(53) IBM BladeCenter QS22. http://www-03.ibm.com/systems/bladecenter/hardware/servers/qs22/ (accessed Sep 17, 2010).

(54) bwGRiD, member of the German D-Grid initiative, funded by the Ministry for Education and Research (Bundesministerium für Bildung und Forschung) and the Ministry for Science, Research and Arts Baden-Wuerttemberg (Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg). http://www.bw-grid.de (accessed Sep 17, 2010).

CT100533U

JCTC Journal of Chemical Theory and Computation

Trends in Aromatic Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes: A Valence Bond Modeling

Sason Shaik,* Petr Milko, Patric Schyman, Dandamudi Usharani, and Hui Chen

The Institute of Chemistry and the Lise Meitner-Minerva Center for Computational Quantum Chemistry, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

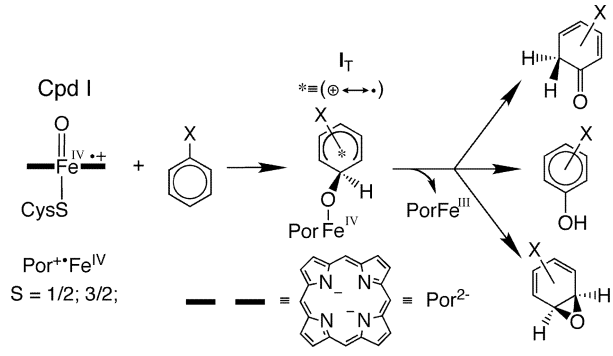
Received September 25, 2010

Abstract: The mixed density functional theory (DFT) and valence bond study described herein focuses on the activation of 17 benzene derivatives by the active species of Cytochrome P450, so-called Compound I (Cpd I), as well as by the methoxy radical, as a potentially simple model of Cpd I (Jones, J. P.; Mysinger, M.; Korzekwa, K. R. *Drug Metab. Dispos.* **2002**, *30*, 7−12). Valence bond modeling is employed to rationalize the P450 mechanism and its spin-state selectivity from first principles of electronic structure and to predict activation energies independently, using easily accessible properties of the reactants: the singlet−triplet excitation energies, the ionization potentials of the aromatics, and the electron affinity of Cpd I and/or of the methoxy radical. It is shown that the valence bond model rationalizes all the mechanistic aspects and predicts activation barriers (for 35 reactions) with reasonable accuracy compared to the DFT barriers with an average deviation of ±1.0 kcal·mol^{-1} (for DFT barriers, see: Bathelt, C. M.; Ridder, L.; Mulholland, A. J.; Harvey, J. N. *Org. Biomol. Chem.* **2004**, *2*, 2998−3005). The valence bond modeling also reveals the mechanistic similarities between the P450 Cpd I and methoxy reactions and enables one to make predictions of barriers for reactions from other studies.

Introduction

Cytochrome P450s are heme enzymes that metabolize and biosynthesize essential compounds,[1] by use of a high-valent iron-oxo porphyrin cation−radical complex, Por$^{+\bullet}$Fe(IV)O, so-called Compound I (Cpd I).[1−3] Among these reactions are the activation of aromatic compounds, Scheme 1, to arene oxides, phenols, and ketones, which influences the bioavailability of drugs (phenols) and may also contribute to carcinogenicity via DNA mutations (arene oxides).[4] The relationship between the various products became intriguing when mechanistic investigations[4b,c] led to the conclusion that the arene oxide is an obligatory intermediate in this reaction and the phenol and ketone are its byproducts.[4d,e] However, new evidence suggested alternative pathways proceeding through radical and/or cationic Meisenheimer tetrahedral intermediates, as shown in Scheme 1.[1a,2e,5] These mechanistic studies have also generated many relative reactivity and regioselectivity data,[2e,5,6] which were addressed by a few groups[5,7,8] and reviewed.

Scheme 1. Intermediates and Products during Arene Oxidation by P450 Cpd I



The advent of density functional theory (DFT) has enabled testing of these mechanistic alternatives on model systems[9] and within native proteins (CYP 2C9)[10] using density functional theory/molecular mechanics (DFT/MM) calcula-

* Corresponding author. E-mail: sason@yfaat.ch.huji.ac.il. Telephone: +972-2-6585909.

10.1021/ct100554g
Published on Web 01/05/2011

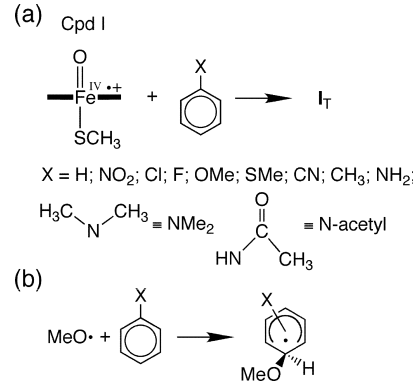
tions. Thus, all the calculations support the intermediacy of a Meisenheimer tetrahedral intermediate (I_T),[3e,9,10] as the source of all products, and reveal the π-activation step as rate determining for all the products. Additionally, the DFT calculations have produced a wealth of information on the dependence of these rate-controlling barriers on the aromatic ring substituents and on the positional selectivity. By and large, these model studies reproduced the observed experimental trends.[9b−e]

In addition the calculations revealed several intriguing features, which we address herein: (i) unlike alkane hydroxylation, which proceeds via the two spin states of Cpd I (S = 1/2; 3/2),[3] aromatic activation preferentially takes place via the lower-energy doublet state (S = 1/2);[9] (ii) the transition states were found to have a hybrid nature with radical and cationic characters;[9,10] (iii) this hybrid character is retained in the tetrahedral intermediates, which is neither fully cationic nor radicalar;[10] (iv) relative to benzene, both electron-donating and -withdrawing substituents decrease the barrier for para position attack;[6a,9c] and (v) in accord with experiment,[6a] a significant preference is observed for para regioselectivity even with electron-withdrawing substituents, e.g., NO_2,[9c,d] which in electrophilic substitution normally leads to meta regioselectivity.

This abundance of knowledge has created the need for establishing order; namely, the outlining of broad generalizations as well as the creation of more intuitive interfaces between experimental and theoretical data. Several studies were published, which employed the methoxy radical as a Cpd I mimic[8a,b,9d] or used a "hybrid" Hammett substituent parameter,[9c] to describe reactivity of Cpd I with aromatic substrates. In the present study, we use valence bond (VB) modeling of aromatic oxidation by P450, with an aim of deriving the above trends from first principles and thereby generating a general theoretical framework that organizes the reactivity patterns. The VB diagram model was previously applied successfully to address reactivity patterns in alkane hydroxylation and thioether sulfoxidation by P450.[11] The Manchester group[12] has extended the VB modeling to include also bond activation by nonheme oxo-iron reagents. The VB diagram model[13] has a few merits: It reveals the origins of the barrier, describes the formation of transition states and reaction intermediates, and allows the prediction of barrier heights and structure−reactivity relationships. As shall be shown, the modified application of the VB diagram model used herein enabled us to go beyond previous treatments[11,12] and derive the above reactivity patterns from first principles based on physically clear predictors. Thus, the modified VB model rationalizes the hybrid nature of the transition states and intermediates[9c,d] as well as the different barriers of the spin states during the reactions with Cpd I and the relationship to radical attacks by MeO•.[8a] Furthermore, this VB model leads to expressions that estimate barrier heights from easily accessible reactant properties, such as singlet−triplet promotion energies, ionization potential (IP), and electron affinity (EA).

Our focus is the series of reactions in Scheme 2a, studied before by Bathelt et al.[9c] using DFT (B3LYP) calculations. As noted by the authors,[9c] some of these molecules would undergo preferentially other reactions and were used by them for the sole purpose of modeling structure−reactivity relationships. Our goal herein is the same. Note that unlike our usual choice to represent the cysteinate axial ligand by HS^-,[3a,b,d,e] we use here CH_3S^- in keeping with the original study of Bathelt et al.[9c] Furthermore, since, Bathelt et al.[9c] showed that the effect of bulk polarity makes a contribution to the barrier, which is virtually substituent independent, and DFT/MM calculations of benzene activation by P450 2C9, confirmed this incremental contribution of bulk polarity to the barrier,[10] we restrict our study to the gas-phase model reactions.[14] To test the reactivity patterns of aromatic activation by a simple radical, we use the reaction series with MeO• in Scheme 2b. The so derived VB insight will be demonstrated by attempts to predict trends in other molecules.

Scheme 2. Studied Reactions of Ar-X Molecules with (a) Cpd I and (b) MeO•



Methods

Software. The starting points for the calculations of the Cpd I addition were the structures published in study of Bathelt et al.[9c] which employed Jaguar 4.2.[15a] Single-point calculations of the barriers for π-activation were carried out with Gaussian 03 and 09,[15b,c] at these structures (communicated by Harvey). In two cases (X = NMe_2, Cl), the calculated energies did not correspond to those obtained in the original study, but upon reoptimization of the TSs with Jaguar 7.6,[15d] the correct structures were obtained, as shown by the new calculated barriers matching the original study. As such, we were able to create a data set wherein all barriers are calculated using the same methods and procedures, thus removing nonsystematic errors, which might be contributed by use of different software packages and procedures (see Supporting Information, Tables S1 and S2).

We note here that in the original study,[9c] the authors optimized Cpd I(CH_3S^-) within C_s symmetry constraints. Removing this constraint and optimizing at C_1 lowers the energy of Cpd I(CH_3S^-) by 4.0 kcal·mol^{-1} (see Supporting Information, Table S3; note that the electronic structure of the Cpd I(CH_3S^-) shows more pronounced mixing of the sulfur p_z orbital with the porphyrin a_{2u} orbital in C_s symmetry than in C_1 one).[3b] Since this adds a constant to the barriers and it had no effect on the quality of the VB modeling, we present here the barriers with the C_s constrained Cpd I(CH_3S^-) to stay as close as possible to the original study.[9c]

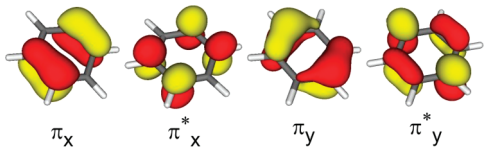


Figure 1. The π and π^*-type orbitals of benzene.

The VB modeling with the C_1 data is given in the Supporting Information (Table S17).

Since we will later attempt to make predictions on cases calculated with a Cpd I model having a HS^- axial ligand, we tested the ligand effect (HS^- vs CH_3S^-) on the benzene activation, using the same basis sets as Bathelt et al.[9c] We found that the axial ligand effect on the calculated benzene activation is small (0.2 kcal·mol^{-1}) if Cpd I(CH_3S^-) is indeed constrained to C_s symmetry. Because of this constraint, we did not perform zero point energy (ZPE) correction, which is anyway small for this kind of reaction.

All the MeO$^\bullet$ reactions as well as IP's and EA's were studied using Gaussian 09 geometry optimization.[15c] Charge-transfer values (see Figure 3) in the transition state were calculated with NBO 3.1 as implemented in Gaussian 03.[15c,16]

Functional and Basis Sets. Thus, as in the original study,[9c] all the calculations were performed using the unrestricted hybrid density functional method UB3LYP.[17] Geometry optimizations (without constraints) were performed with the LACV3P basis set on iron and 6-31G* on the rest of the atoms (basis set BSI).[17,18,19a,b] Subsequently, single point calculations were done on the optimized geometries using BSII, which corresponds to LACV3P(Fe)/6-311+G** (rest).[19c,d] The so computed reaction barriers for the Cpd I/arene series (Scheme 2a) were within ±0.6 kcal·mol^{-1} of those reported in the study of Bathelt et al. with an exception for addition to the meta position in which the deviation was 1.3 kcal·mol^{-1} (see Supporting Information, Tables S1 and S2).[9c] The MeO$^\bullet$ transition states were optimized with the 6-31G* basis set, and energy was corrected using the 6-311+G** basis set.

Auxiliary Data for VB Modeling. As shall be seen, the VB modeling relies on two properties of the arene molecules, the vertical IP and singlet−triplet $\pi\pi^*$ excitation energies, ΔE_{ST}. Further, it requires also EA's of Cpd I and of the methoxy radical as input data. We used the experimental IP values of the studied substrates from the NIST database.[20] In parallel we ascertained that DFT reproduces these IP's well (see Supporting Information for details).

To be consistent with calculated IP's (see Supporting Information, Figure S1 and Table S7), the B3PW91/6-311++G** level was chosen to obtain the vertical ΔE_{ST} values.[17a,19c,d,21] Based on the π-type orbitals in Figure 1, there are generally two closely lying excitation types, which can be obtained from DFT and involve π_y to π^*_y and π_x to π^*_x excitations. The $\pi_x \rightarrow \pi^*_x$ excitation energy is insensitive to the nature of the substituents, while the $\pi_y \rightarrow \pi^*_y$ excitation is strongly dependent on the substituent, giving generally lower values for the latter excitation. A plot of the calculated $\Delta E_{ST}(\pi_y \rightarrow \pi^*_y)$ excitations against adiabatic experimental values[22] shows identical trends (see Supporting Information,

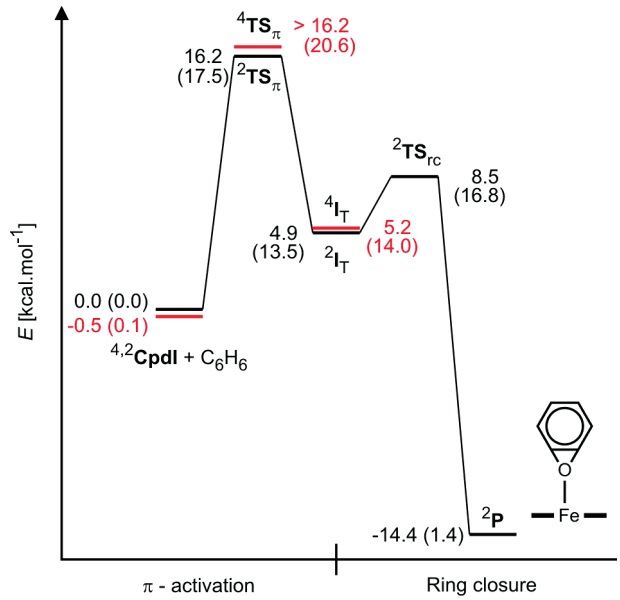


Figure 2. B3LYP potential energy profiles for the epoxidation of benzene by ^{4,2}Cpd I (the quartet-state species are marked in red). All energies are in kcal·mol^{-1} relative to isolated ^{2}Cpd I and benzene. Each species has two energy values, corresponding to BSII (this work and ref 9c) and in parentheses to LACVP(Fe)/6-31G(rest) from ref 9a.

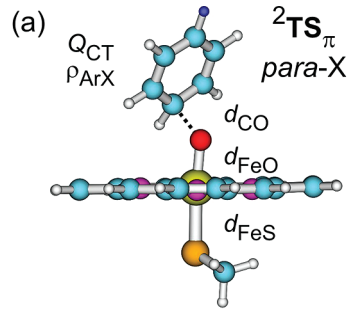
Figure S2), and with the exception of X = NO_2, the calculated vertical values are about 20 ± 3 kcal·mol^{-1} higher than the experimental adiabatic values. The calculated $\Delta E_{ST}(\pi_y \rightarrow \pi^*_y)$ value for benzene 102.2 kcal·mol^{-1} is in excellent agreement with the CCSD(T)/cc-pV∞Z calculated datum, 104.4 kcal·mol^{-1}, for the $^3B_{1u}$ state[23a] and close to a spin-coupled valence bond (SCVB) calculated value, 97.3 kcal·mol^{-1}.[23b] From absorption peak progressions for benzene and fluorobenzene in magnetic induced singlet−triplet excitations studied by Evans,[22b] it is possible to deduce that the vertical excitation energies are ∼94−97 kcal·mol^{-1} compared with the DFT calculated values 102.2 and 101.9 kcal·mol^{-1}.

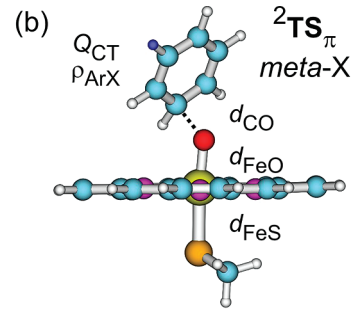
The EAs of Cpd I and MeO$^\bullet$ are constant quantities for the respective reaction series studied here, but to be consistent with past calculations,[3a,b,d,e] B3LYP was used to obtain the vertical EA of Cpd I using C_s geometry,[9c] leading to EA = 64.9 kcal·mol^{-1}. The unconstrained C_1 structure has a lower EA value of 60.6 kcal·mol^{-1}, while the Cpd I with HS^- ligand has EA = 67.9 kcal·mol^{-1}. The vertical EA of MeO$^\bullet$ was determined using single point calculations at CCSD(T) level of theory,[24] CCSD(T)/aug-cc-pVQZ//UB3LYP/6-311+G**, and leads to EA = 32.1 kcal·mol^{-1}.[25]

All the data generated in this study are shown in the Supporting Information document. For space economy, the following sections will focus on the key data only.

Results

Energy Profiles. Figure 2 shows the energy profiles for the activation of benzene by Cpd I, using the recalculated data based on geometries from Bathelt at el.[9c] and the previous data of de Visser and Shaik using Cpd I(HS^-).[9a] Despite the differences in the representations of Cpd I and





X	d_{CO}	d_{FeO}	d_{FeS}	Q_{CT}	ρ_{ArX}
H	1.890	1.730	2.378	0.35	-0.46
Cl	1.902	1.731	2.361	0.34	-0.44
F	1.950	1.717	2.386	0.33	-0.38
CN	1.939	1.719	2.369	0.24	-0.34
NO_2	1.847	1.727	2.411	0.21	-0.41
NMe_2	2.052	1.693	2.389	0.46	-0.45
OMe	2.006	1.710	2.369	0.40	-0.40
NH_2	2.077	1.693	2.384	0.43	-0.41
CH_3	2.028	1.701	2.396	0.32	-0.35
SMe	1.951	1.711	2.389	0.38	-0.44
N-acetyl	2.064	1.701	2.391	0.33	-0.35

X	d_{CO}	d_{FeO}	d_{FeS}	Q_{CT}	ρ_{ArX}
Cl	1.913	1.720	2.383	0.28	-0.37
F	1.866	1.733	2.376	0.32	-0.46
CN	1.905	1.733	2.340	0.26	-0.39
NO_2	1.884	1.733	2.388	0.25	-0.38
NMe_2	1.872	1.725	2.376	0.41	-0.48
OMe	1.974	1.711	2.380	0.34	-0.39

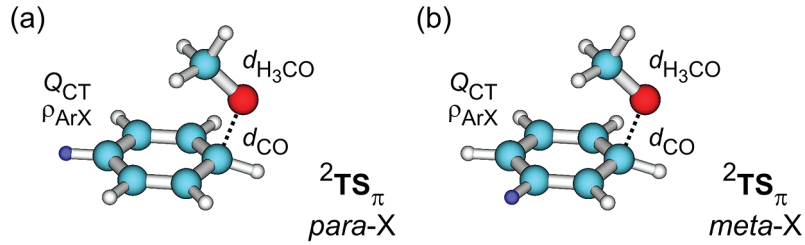
Figure 3. Optimized $^2TS_\pi$ species during π-activation of arenes by Cpd I, key geometric distances (in Å), degrees of charge transfer (Q_{CT}) from the arene to Cpd I, and spin densities on the arenes (ρ_{ArX}). TSs for (a) para attack and (b) meta attack. Color code: yellow, orange, red, purple, and blue correspond to iron, sulfur, oxygen, nitrogen, and carbon, respectively. Dark blue represents the position of the X substituents.

the basis sets, the two sets of relative energy values are mechanistically consistent. Thus, the initial step involves the π-activation of benzene by ^{4,2}Cpd I via two transition states, $^{4,2}TS_\pi$. In both studies, the quartet species $^4TS_\pi$ lies about 0.6−3.3 kcal·mol^{-1} above the corresponding doublet transition state. Recalculating the barrier with Cpd I(HS^-) and the same basis sets as those in Bathelt et al.[9c] gave a 1.8 kcal·mol^{-1} preference for the doublet over the quartet state. In all cases, these transition states lead to the corresponding tetrahedral Meisenheimer intermediates ($^{4,2}I_T$), and again the doublet-state species is lower in energy.[9a−c] The intermediate in turn undergoes a variety of reactions (see Scheme 1), and Figure 2 shows the ring closure to the benzene-oxide, via $^2TS_{rc}$, which represents the simplest reaction pathway toward a product. As found by de Visser and Shaik,[9a] the quartet-state profile continues to lie above the doublet state. In both studies the doublet-state barrier to ring closure is smaller than those for the π-activation step. All other barriers for the conversion of ^2I_T to the other two products (phenol and ketone, in Scheme 1) are also small.[9a,c] Therefore, the VB modeling will focus hereafter on the π-activation step in the doublet spin state. Since none of the follow-up steps is rate controlling, their VB modeling will be largely waved (with the exception of the qualitative representation in Figure 7a of the simplest follow-up step).

Transition States for π-Activation by Cpd I and MeO$^\bullet$. Key geometric features of the $^2TS_\pi$ species for π-activation of the various substituted benzene derivatives, studied herein, are shown in Figure 3. Figure 4 displays the corresponding species for the reactions with the MeO$^\bullet$ radical (Scheme 2b), which focused more on para attacks. For each TS_π in Figures 3 and 4, we indicated three bond distances and the quantities: Q_{CT}, the amount of charge transferred (CT) from the arene to Cpd I in the TS, and ρ_{ArX}, the spin density value on the ArX molecule in the TS.

Inspection of Figure 3 shows that: (i) all the $^2TS_\pi$ species are uniformly side-on types; this uniformity is important since side-on and face-on barriers have small differences, which would have added nonsystematic contributions into the data set; (ii) all the $^2TS_\pi$s possess a hybrid radical/cationic character in the arene; (iii) the O···C bond lengths in the transition states vary in the range of 1.847−2.077 Å; (iv) the $^2TS_\pi$ species with para electron-releasing substituents have 'earlier' structures with longer C···O bond lengths; (v) the para substituted $^2TS_\pi$ species are significantly earlier than the meta substituted ones; and (vi) the amount of charge transferred from the arene to Cpd I, Q_{CT}, depends on the substituent; it is larger for the electron-releasing substituents with maximum of Q_{CT} = 0.46e for NMe_2, and the effect is more significant for the para substituted $^2TS_\pi$ species. Figure 4 shows similar trends in Q_{CT} values and O···C distances but within a narrower range compared with the Cpd I reactions.

Barriers for π-Activation by Cpd I and MeO$^\bullet$ and Trends. Table 1 collects the π-activation barriers for the P450 reactions as well as those calculated for the reactions with MeO$^\bullet$, along with two properties of the arene: the IP's and the singlet−triplet $\pi\pi^*$ excitations, $\Delta E_{ST}(\pi\pi^*)$.



X	d_{H_3CO}	d_{CO}	Q_{CT}	ρ_{ArX}
H	1.396	1.898	0.23	0.49
Cl	1.396	1.908	0.23	0.48
F	1.396	1.899	0.24	0.49
CN	1.395	1.920	0.19	0.45
NO_2	1.396	1.911	0.18	0.44
NMe_2	1.390	1.940	0.30	0.53
OMe	1.393	1.916	0.27	0.51
NH_2	1.392	1.935	0.28	0.51
CH_3	1.394	1.910	0.25	0.49
SMe	1.393	1.933	0.26	0.50
N-acetyl	1.393	1.927	0.25	0.49

X	d_{H_3CO}	d_{CO}	Q_{CT}	ρ_{ArX}
NO_2	1.399	1.881	0.20	0.47
NMe_2	1.393	1.891	0.26	0.53

Figure 4. Optimized $^2TS_\pi$ species during π-activation of arenes by MeO$^\bullet$, key geometric distances (in Å), degrees of charge transfer (Q_{CT}) from the arene to Cpd I, and spin densities on the arenes (ρ_{ArX}). TSs for (a) para attack and meta attack. Color code: red and blue correspond to oxygen and carbon, respectively. Dark blue represents the position of the X substituents.

Table 1. Experimental IP's, Calculated $\pi\pi^*$ Singlet−Triplet Excitations, $\Delta E_{ST}(\pi\pi^*)$, and UB3LYP/BSII/UB3LYP/BSI Calculated Barriers, $\Delta E^{\ddagger}$ (kcal·mol^{-1}), for π-Activation of Ar-X by Cpd I and MeO$^\bullet$ on para and meta Positions

X	IP_{exp}[a] (kcal·mol^{-1})	ΔE_{ST} ($\pi\pi^*$)[b] (kcal·mol^{-1})	$\Delta E^{\ddagger}$, CpdI (kcal·mol^{-1})[c]		$\Delta E^{\ddagger}$, MeO$^\bullet$ (kcal·mol^{-1})[c]	
			para	meta	para	meta
H	213.1	102.2	16.2		7.9	
Cl	209.4	97.3	15.3	15.7	7.4	
F	214.0	101.9	15.2	16.3	7.5	
CN	225.3	89.3	14.9	15.9	7.6	
NO_2	232.0	87.1	14.2	15.3	8.1	9.0
NMe_2	174.1	89.4	9.6	16.8	3.3	8.1
OMe	193.7	97.6	13.2	15.3	5.7	
NH_2	185.4	91.5	11.0		4.2	
CH_3	205.0	99.0	15.0		7.0	
SMe	187.7	89.5	12.6		5.3	
N-acetyl	195.1	90.7	13.6		5.9	

[a] Experimental values from the NIST database. [b] Calculated values (B3PW91/6-311++G**). [c] Without ZPE correction.

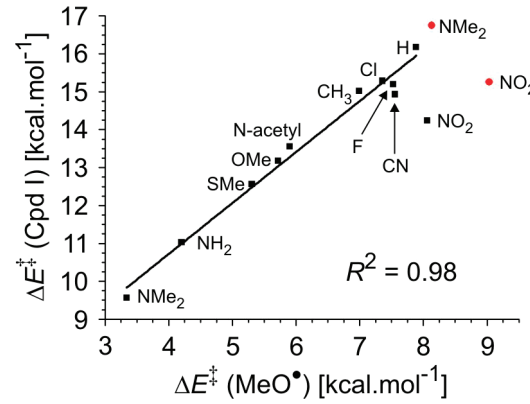


Figure 5. A plot of the barriers for π-activation of arenes by Cpd I vs MeO$^\bullet$. The red circles are barriers for meta-position attacks (without ZPE correction). The R^2 value corresponds to the points of the best fit for the attack on the para position (excluding the NO_2−Ph data).

Inspection of Table 1 shows the following trends: (i) The P450 barriers are sensitive to the substituent on the benzene ring and vary between 9.6 kcal·mol^{-1} (Ph-NMe_2) to 16.2 kcal·mol^{-1} (benzene); (ii) the barriers for attack on the meta positions are generally larger than those for the para position in which the most pronounced effect is observed for the Cpd I addition to Ph-NMe_2 (9.6 and 16.8 kcal·mol^{-1} for the addition on para and meta position, respectively); (iii) the barriers for MeO$^\bullet$ para-position attacks are much smaller than the P450 values and vary in a narrower range of 3.3−8.1 kcal·mol^{-1}; and (iv) with the exception of X = NO_2 in the MeO$^\bullet$ series, in both P450 and MeO$^\bullet$ series, the attacks on para positions of the substituents have lower barriers relative to benzene. Indeed, activation barriers of the Cpd I addition and the methoxy radical addition are in good mutual correlation, with the exception of nitrobenzene (Figure 5).

The $\Delta E^{\ddagger}$ values for the para-position attack (by either Cpd I or MeO$^\bullet$), with the exception of those for the most electron-withdrawing substituents (CN, NO_2), show a linear dependence on the IP values of the arene.[12d] Similarly, part of the data correlates linearly with the $\Delta E_{ST}(\pi\pi^*)$ values. However, none of these two physical properties can by itself correlate with all of the data. By contrast, all the P450 barrier set can be correlated nicely with a hybrid quantity, IP +

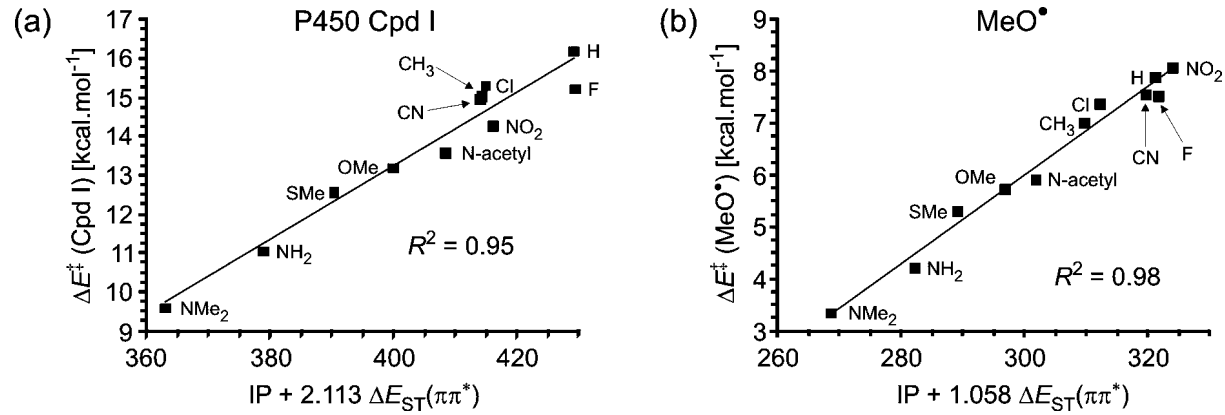


Figure 6. A plot of the π-activation barriers vs: (a) IP + 2.113$\Delta E_{ST}(\pi\pi^*)$, for the P450 reactions and (b) IP + 1.058$\Delta E_{ST}(\pi\pi^*)$ for the MeO$^{\bullet}$ reactions. (IP's taken from NIST, and the fit is based on Maple 13 software).

2.113$\Delta E_{ST}(\pi\pi^*)$, as seen in Figure 6a, and for the MeO$^{\bullet}$ set with IP + 1.058$\Delta E_{ST}(\pi\pi^*)$, as shown in Figure 6b. The double correlation was obtained by a standard fit routine, as is implemented in Maple 13 program package (see Supporting Information, Table S9). This hybrid correlation retrieves the similar one found by Bathelt et al.,[9c] using Hammett substituent parameters. The correlation follows also the hybrid character seen in the charge transfer and spin density in the P450 transition states in Figure 3.

Discussion

The above computational results show a few intriguing trends for the P450 reactions:[9] (i) The computed P450 profiles show that the doublet-state mechanism is lower in energy relative to the quartet state; (ii) the π-activation step has a higher lying barrier than the following rearrangement steps; (iii) the $^2TS_{\pi}$ species for π-activation as well as the corresponding tetrahedral intermediates, ^2I_T, have hybrid radical/cationic characters (Figure 3), which depend on the ring substituent; (iv) the π-activation barriers are sensitive to the substituent on the benzene ring, the lowest barrier is obtained for the Cpd I addition to the para position of Ph-NMe$_2$, but all the para substituents are found to lower the barrier with respect to the unsubstituted benzene; (v) the barriers for attack on the meta positions are generally larger than those for the para position, even for the electronic-withdrawing substituents which generally direct electrophilic reagents for meta attacks; (vi) the π-activation barriers can be correlated reasonably well with a mixed quantity made from a combination of the IP of the arene and its singlet−triplet excitation energy, $\Delta E_{ST}(\pi\pi^*)$; and (vii) the barriers for MeO$^{\bullet}$ attacks are much smaller than the P450 values and vary in a narrower range of 3.3−8.1 kcal·mol^{-1}, nevertheless the two barrier sets exhibit a reasonable mutual correlation.

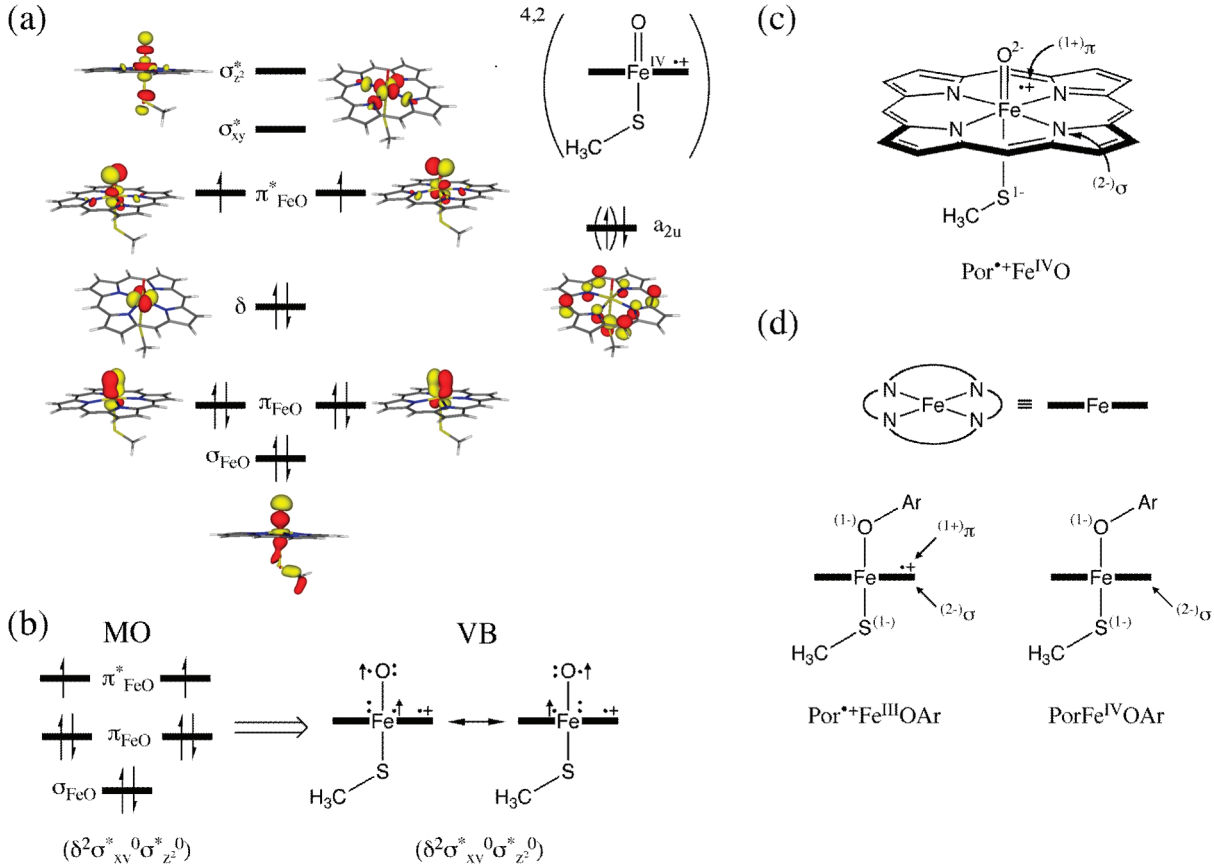
We shall now present a VB modeling of these reactions with an aim of unifying these findings and understanding thereby these reactivity patterns.[13] Subsequently we shall show that the barriers can be calculated from raw data based on the VB model.[11]

VB Modeling. *Energy Profiles Generated from VB Mixing Principles.* Since the follow-up rearrangements of the tetrahedral intermediate have much smaller barriers compared with the common π-activation step, the modeling will focus on this step up to the Meisenheimer complex intermediate. To facilitate the discussion, we show in Scheme 3, the molecular orbital (MO) and VB representations of Cpd I and some helpful oxidation-state formulations.[11] Scheme 3a shows key MOs of Cpd I: The $\sigma_{FeO}^2\pi_{FeO}^4\pi^*_{FeO}{}^2$ configuration represents the bonding block and accounts for a σ_{FeO} bond and a $\pi_{FeO}^4\pi^*_{FeO}{}^2$ manifold, as in $^3O_2$,[3a,b,d,e,26,27] and spin-up electrons in π^*_{FeO}. The π^*_{FeO} orbitals are considered as "d" orbitals, so that the d-block occupancy is $\delta^2\pi^*_{FeO}{}^2\sigma^*_{xy}{}^0\sigma^*_{z^2}{}^0$. Finally, the porphyrin cation−radical is represented by the singly occupied a_{2u}; the double-headed arrow represents spin-down/spin-up arrangements for the doublet/quartet states of Cpd I.

Scheme 3b outlines the correspondence of the MO and VB representations of Cpd I, with the $\delta^2\sigma^*_{xy}{}^0\sigma^*_{z^2}{}^0$ block placed in parentheses. On the left side, we show the bonding block MO configuration, $\sigma_{FeO}^2\pi_{FeO}^4\pi^*_{FeO}{}^2$. In the VB representation σ_{FeO} is drawn as a line, while in VB the π-block is represented by two resonating three-electron bonds, which span two perpendicular planes, with two spin-up electrons on iron-oxo.[3,11] Finally, the open-shell porphyrin (a_{2u}^1) is represented by a cation−radical symbol on porphyrin. These VB cartoons will be used hereafter. The reader may note also that each of the resonance structures, in Scheme 3b, looks like Fe^{III}−O$^{\bullet}$. Nevertheless, because their superposition relays four of the electrons to π_{FeO}-bonding orbitals, this leaves a Fe(d^4) configuration that qualifies as Fe^{IV}O. However, during the reaction the electronic structure gets localized and becomes Fe^{III}−O$^{\bullet}$.[11c]

Scheme 3c and 3d summarizes some basic conventions of the oxidation-state formalism, which tracks d-electron counts of transition-metal complexes during redox processes. Scheme 3c shows oxidation numbers for Cpd I: the porphyrin has a σ-oxidation number of 2−, the oxo is 2−, and the thiolate is 1−. Since the molecule is neutral, the heme oxidation-state is V, which becomes Por$^{+\bullet}$Fe(IV)O, based on spectroscopic evidence[1,2] for a porphyrin π-cation radical. With Fe(IV), Cpd I will have a d^4 electronic configuration.[3,9] Scheme 3d depicts the iron-aryloxo electromers due to π-attack by Cpd I on the arene. The resulting OAr group has an oxidation number of 1−, and hence the effective oxidation state of the heme becomes IV, which can manifest

Scheme 3. MO and VB Representations of Cpd I[a]



[a] The following represent: (a) Key MO's, (b) MO-VB correspondence of the FeO-bonding block, (c) oxidation numbers in Cpd I, and (d) oxidation numbers in tetrahedral intermediate (I_T).

as PorFe(IV)OAr and/or $Por^{+\bullet}Fe(III)OAr$, with electronic configurations d^4 and/or d^5, respectively. As amply discussed,[3] π-activation by Cpd I leads to two electromeric states for the tetrahedral intermediates,[3e,9a] of the $Por^{+\bullet}Fe^{III}OAr$ and $PorFe^{IV}OAr$ types. Since the VB diagrams are very similar for the two electromers[11a] and the latter are usually the more stable in the gas phase, as found to be so in this study, we shall focus only on the latter type.

Figure 7a shows the VB diagram for benzene epoxidation via the doublet spin state. For the sake of economy, the benzene is symbolized by a single Kekulé structure and so is Cpd I. The diagram involves two principal curves for the direct O transfer to the arene. The curves are anchored at the ground states (Ψ_r, Ψ_p) and their two promoted states (Ψ_p*, $\Psi_{r,CT}$*). This direct process is, however, catalyzed by an intermediate-state curve (Ψ_I*(IV)) that cuts through the higher energy ridge for direct oxo-transfer and splits the process into side-on π-attack followed by ring closure to form the ferric-arene-oxide product.[11,13b,c] This three-curve VB diagram is a typical case,[11−13] wherein an intermediate state internally facilitates the otherwise more difficult transformation of Ψ_r directly to Ψ_p.

Let us elaborate on the electronic structure of the promoted state for the principal curves: $^2\Psi_{r,CT}$* is a state with a mixed CT and covalent structures, which describes the two new O−C bonds that will be formed between the oxo of Cpd I and the arene molecule. To save space,[3e,11] Figure 7a shows only the main charge-transfer structure, whereas Figure 7b shows explicitly the contributing structures, which combine together to produce eventually the arene oxide in $^2\Psi_p$. There are two equivalent charge-transfer structures, $^2\Phi_{CT}$, which arise by one-electron transfer from the arene to porphyrin$^{+\bullet}$ and where the electrons on the $O^\bullet$ and $C^\bullet$ are coupled to a bond pair across one C−O linkage, while the other linkage has an ionic character (shown by dots). In addition, there is a purely covalent contributor, $^2\Phi_{COV}$, which maintains two covalent $C^\bullet$—$^\bullet O$ spin pairs between the arene and the oxo of Cpd I. The charge-transfer structures, $^2\Phi_{CT}$, dominate $^2\Psi_{r,CT}$*, as indicated in Figure 7a.[28,29] Note that, since the $C^\bullet$—$^\bullet O$ bond pairing lowers the oxidation number of the heme to a ferric (Fe^{III}) state, $^2\Psi_{r,CT}$* is actually an image state of the ferric-arene oxide product and hence along the reaction coordinate $^2\Psi_{r,CT}$* correlates to $^2\Psi_p$, as shown in Figure 7a. In an analogous manner, in the reverse direction, the promoted state, $^2\Psi_p$*, is formed from $^2\Psi_p$ by an electron transfer from the porphyrin to one of the O−C bonds while pairing the electrons on the arene, and as such, $^2\Psi_p$* is the electronic image of the ground state on the other side, $^2\Psi_r$, and the two correlate along the reaction coordinate.

Let us turn now to the intermediate-state curve in Figure 7a. This VB curve which is anchored in $^2\Psi_I$*(IV) participates in the π-activation step by forming the Meisenheimer intermediate ^2I_T. Thus, Ψ_I*(IV) involves an electron shift from iron (from π_{FeO}*, see Scheme 3a) to porphyrin$^{+\bullet}$ (which

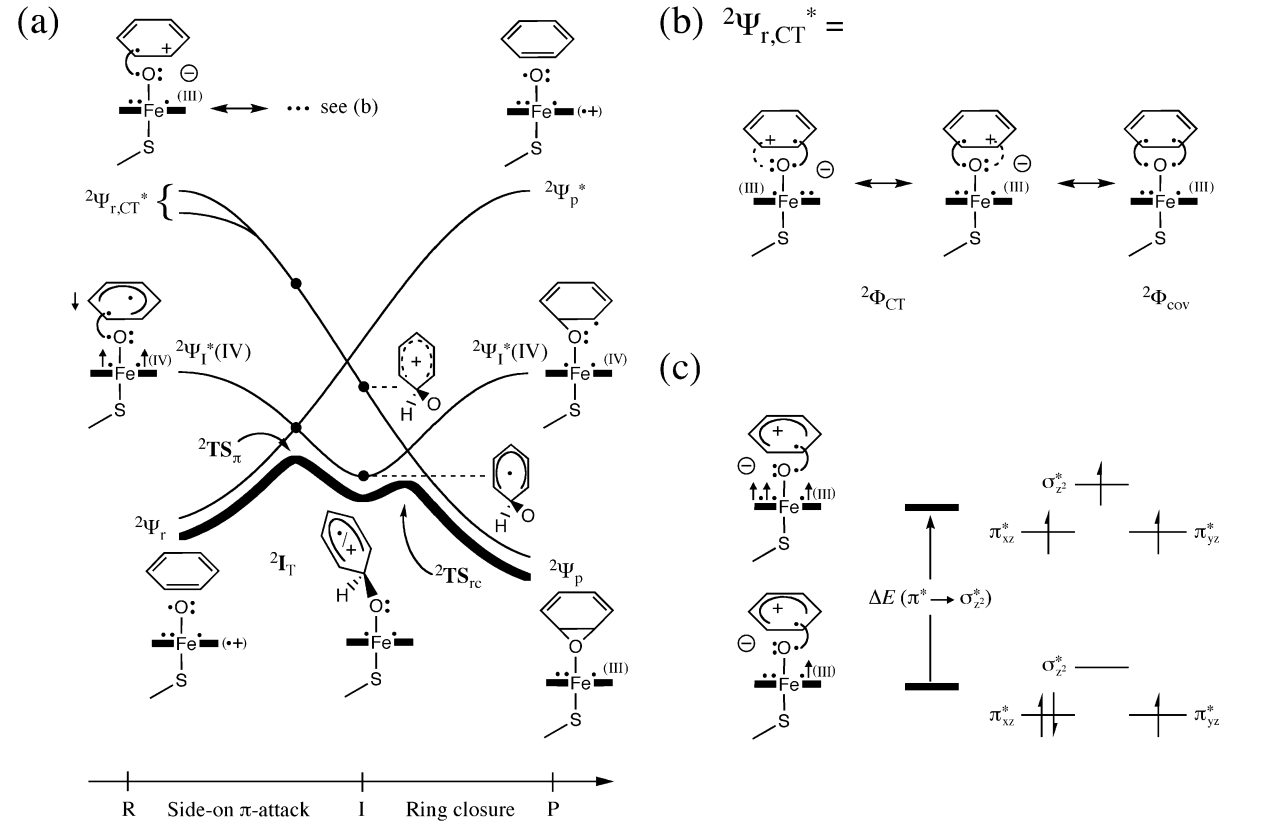


Figure 7. (a) A VB diagram describing the mechanistic scenario during the doublet spin-state conversion of benzene to benzene-oxide via intermediacy of the Meisenheimer intermediate, ^2I_T. (b) A detailed description of the major contributors to the promoted state $^2\Psi_{r,CT}$*. (c) The relative energies of $^2\Psi_{r,CT}$* and $^4\Psi_{r,CT}$*.

is energetically a small excitation, ca. 5−6 kcal·mol^{-1}).[11c] In addition, the π-system of benzene is promoted to a triplet configuration, while C$^\bullet$ and $^\bullet$O are coupled into a bond pair, which eventually becomes the O−C bond in ^2I_T. Note that $^2\Psi_I$*(IV) is different than the covalent component $^2\Phi_{COV}$ of $^2\Psi_{r,CT}$*, and it also lies lower in energy, since $^2\Phi_{COV}$ involves a costly promotion from the oxygen doubly occupied orbital to porphyrin$^{+\bullet}$ (π_{FeO} → a_{2u}).[28] Thus, $^2\Phi_{COV}$ possesses two C$^\bullet$—$^\bullet$O bond pairs, while $^2\Psi_I$*(IV), wherein the oxo group has three electrons, can form only one C$^\bullet$—$^\bullet$O bond pair, and therefore the latter leads to the Meisenheimer intermediate. On the product side, $^2\Psi_I$*(IV) correlates to an excited state of the product having Fe(IV) and three-electrons in one of the C−O linkages.

The final energy profile in Figure 7a is obtained by the mixing of the three state curves, resulting in a biphasic energy profile, dominated by the intermediate-state curve, with a π-activation phase followed by ring closure. The relative barrier heights are determined by the vertical promotion energies between the intersection states, and since the promotion gap at the ^2I_T junction is much smaller than in the reactant onset, the barrier for the π-activation phase is rate controlling, while at the ring-closure state it is a rather small barrier. Other follow-up steps from the ^2I_T junction, e.g., the formation of phenols, can be described analogously, but they require their own VB diagrams,[13] as they are associated with the migration of the ipso proton to the oxo ligand.[9a]

Nature of the π-Activation Transition State and the Tetrahedral Intermediate. Let us start with the natures of $^2TS_\pi$ and ^2I_T Figure 7a. The charge-transfer state curve, $^2\Psi_{r,CT}$* lies not so much higher than the intersection point of the $^2\Psi_r$—$^2\Psi_I$*(IV) curves, where $^2TS_\pi$ will be formed by mixing of the three state curves. Consequently, $^2TS_\pi$ will exhibit a partial charge transfer from the arene to the Cpd I moiety, to an extent that depends on the arene substituent (X). For example, with X = NMe_2 the IP of the arene is the lowest in the series, ca. 58 kcal·mol^{-1} lower than that of nitrobenzene, with the highest IP. As such, the mixing of the charge-transfer state for X = NMe_2 will be the most pronounced in the series, while for X = NO_2, it would be the least significant, as in fact revealed by the computational results in Figure 3, which shows Q_{CT} values for the various substituents. Similarly, the tetrahedral ^2I_T intermediate will have a hybrid character, with a dominant radical character, but neither fully radical nor fully cationic.

Spin-State Preference. The difference between the doublet and quartet spin processes is depicted in Figure 7c, which shows $^2\Psi_{r,CT}$* and $^4\Psi_{r,CT}$* and their energy difference. Thus, in both states we have a PorFe(III)O$^\bullet$—$^\bullet$Ar$^+$ species, which arises by electron transfer from the arene to porphyrin$^{+\bullet}$ while coupling O$^\bullet$ and $^\bullet$C to a bond-pair. However, whereas the PorFe(III) moiety of $^2\Psi_{r,CT}$* has a $\delta^2\pi^{*2}\pi^{*1}$ d-block configuration, $^4\Psi_{r,CT}$* is typified by $\delta^2\pi^{*1}\pi^{*1}\sigma^{*1}_{z^2}$. Thus, $^4\Psi_{r,CT}$* involves also a π^* → $\sigma^*_{z^2}$ promotion within the d-block (approximately, 30 kcal·mol^{-1}).[11b] Since the charge-transfer state is rather close to the other two curves, its mixing into $^{2,4}TS_\pi$ can be deduced from simple considerations of perturbation theory. It is thus expected that generally the $^2TS_\pi$ species will have greater mixing and be lower than $^4TS_\pi$.

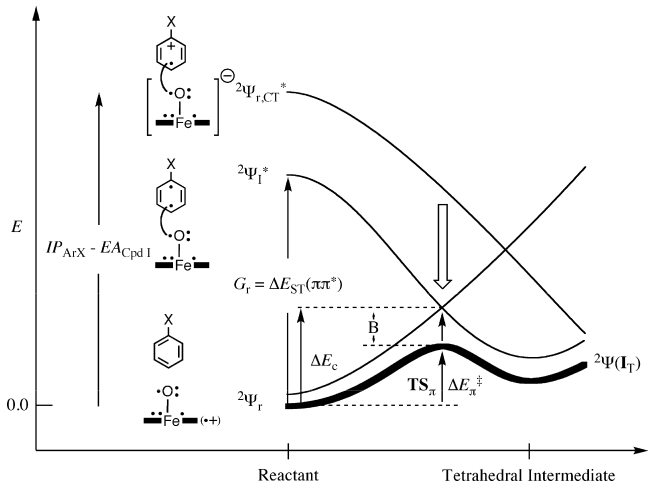


Figure 8. A VB diagram for the π-activation step, showing the three state curves and the key quantities that determine the barrier.

Of course, the extent of this spin selectivity is expected to be modulated by the substituent X, such that electron-withdrawing susbstituents, which raise the IP of the arene, will lead to a smaller energy advantage of $^2TS_\pi$ over $^4TS_\pi$ species.

Reactivity Patterns: Using VB Diagrams to Estimate π-Activation Barriers. Figure 8 shows a section of the VB diagram with the reactivity quantities, which are necessary for modeling of the barrier and its variation. The drawings of the species near the curve represent an attack on the para position to the substituent X, and later on we shall generalize this to meta attacks. The figure shows that the barrier is determined largely by the avoided crossing and VB mixing of the reactant and intermediate curves, $^2\Psi_r$ and $^2\Psi_I{}^*$, but the charge-transfer state lying above, $^2\Psi_{r,CT}{}^*$, can also mix and lower the resulting barrier. The simplest expression for the barrier of an elementary step is eq 1:

$$\Delta E_\pi^\ddagger = \Delta E_c - B \tag{1}$$

Here ΔE_c measures the height of the crossing point of Ψ_r and $^2\Psi_I{}^*$, and B is TS-resonance energy due to the VB mixing of the three curves.

The height of the crossing point reflects the total deformation energies of the two reactants and their Pauli repulsions, which are required to achieve the $^2\Psi_r - {}^2\Psi_I{}^*$ crossing.[30] As usual, the height of the crossing point can be expressed as a fraction (f) of the promotion energy at the reactant side (G_r) leading to eq 2:

$$\Delta E_\pi^\ddagger = fG_r - B \tag{2}$$

Since the promotion energy is simply the singlet-to-triplet excitation of the arene, eq 2 becomes

$$\Delta E_\pi^\ddagger = f\Delta E_{ST} - B \tag{3}$$

Recalling that B reflects also the mixing in the charge-transfer state, we expect that this quantity will vary as a function of the relative energy of the transfer state $\Psi_{r,CT}{}^*$ near the crossing point. This energy difference cannot be quantified computationally, but we should expect that it will change in proportion to the initial energy of $^2\Psi_{r,CT}{}^*$ relative to the ground state, $^2\Psi_r$, and is given by:

$$\Delta E_{CT}(^2\psi_r \rightarrow {}^2\psi_{r,CT}{}^*) = IP_{ArX} - EA_{Cpd\ I} \tag{4}$$

where IP_{ArX} is the IP of the arene, while $EA_{Cpd\ I}$ is the EA of Cpd I.

In summary, we expect that the barrier will be determined by the variation of the singlet-to-triplet excitation of the arene, with a secondary influence of the ionization potential of the arene. Equations 3 and 4 can be used to estimate barriers for the series of reactions of this study. Since the reaction resembles a radical attack, we can use $f = 0.3$ or 1/3, as done previously for H abstraction for radicals.[11a,31] The only missing quantity is then B. However, having f and ΔE_{ST}, we can extract the B values needed to reproduce the DFT-calculated barriers:

$$B = f\Delta E_{ST} - \Delta E_\pi^\ddagger \tag{5}$$

These data are shown in the fifth column of Table 2. Thus, for example, using the DFT-calculated barrier ($\Delta E_\pi^\ddagger$) of benzene, the corresponding ΔE_{ST} value (102.2 kcal·mol^{-1}), and $f = 0.3$, we get B(benzene) = 14.5 kcal·mol^{-1}, and the same procedure leads to $B = 17.2$ kcal·mol^{-1} for the *N*,*N*-dimethylaniline. Other B values are derived similarly.

But we can do better than that, by modeling the B values based on the understanding that these quantities reflect the mixing of the corresponding charge-transfer states. Using perturbation theory, this mixing will be inversely proportional to the energy gap between the charge transfer $\Psi_{r,CT}{}^*$ and the crossing point in Figure 8 and will be proportional to the matrix element that couples the states. Since the energy gap of the crossing point is expected to be proportional to $IP_{ArX} - EA_{Cpd\ I}$ in eq 4 and the matrix element for coupling these states is gauged by the odd electron density on the carbon site where O−C bond is made, we can use the following simple expression for B_X for a given substituent X, relative to B_H for the unsubstituted benzene:

$$B_X = B_H\{[\rho_X(IP_H - EA_{Cpd\ I})]/[\rho_H(IP_X - EA_{Cpd\ I})]\} \tag{6}$$

Here ρ_X and ρ_H are, respectively, the spin densities at the sites of attack of the X-substituted arene vs benzene, in the corresponding triplet states. Thus, all we need is to have B_H, the value for benzene, and derive from it all the other B_X values. The so calculated B values are collected in the sixth column of Table 2.

Before looking at these results it is instructive to inspect the spin densities, which are depicted in Figure 9 for representative substituents (for remaining ones see Supporting Information, Figures S3 and S4). For most of the substituents, the lowest triplet state is due to $\pi_y \rightarrow \pi^*_y$ excitation (consult Figure 1 for the orbitals), and Figure 9a shows the spin densities for three cases. It is apparent that the site of the highest spin density is the para position, whereas the meta position has negligible spin density. Figure 9b shows the triplet spin densities corresponding to the $\pi_x \rightarrow \pi^*_x$ excitation. Here it is seen that the spin density is largest at the meta and ortho positions, but the meta spin

Table 2. Reactivity Factors and VB Estimated B_X Values and Barriers for the para Position Attacks by Cpd I on ArX Molecules

X	IP[a] (kcal·mol^{-1})	ΔE_{ST} ($\pi\pi^*$)[b] (kcal·mol^{-1})	$\rho_{x,para}$[c]	B_X[d] (DFT)	B_X[e] (VB)	$\Delta E_\pi^\ddagger$ (VB)[f] (kcal·mol^{-1})	$\Delta E_\pi^\ddagger$ (DFT) (kcal·mol^{-1})
H	213.1	102.2	0.86	14.5	14.5	16.2	16.2
Cl	209.4	97.3	0.80	13.9	13.8	15.4	15.3
F	214.0	101.9	0.83	15.4	13.9	16.7	15.2
CN	225.3	89.3	0.75	11.9	11.7	15.1	14.9
NO_2	232.0	87.1	0.76	11.9	11.4	14.8	14.2
NMe_2	174.1	89.4	0.76	17.2	17.4	9.4	9.6
OMe	193.7	97.6	0.82	16.1	15.9	13.4	13.2
NH_2	185.4	91.5	0.76	16.4	15.7	11.7	11.0
CH_3	205.0	99.0	0.89	14.7	15.9	13.8	15.0
SMe	187.7	89.5	0.75	16.2	15.2	13.5	12.6
N-acetyl	195.1	90.7	0.75	13.8	14.4	13.0	13.6

[a] Experimental values from the NIST database corresponding to vertical ionization. [b] See Table 1. [c] Spin density localized at the para carbon of ArX in the triplet state. [d] B_X is defined by eq 5; $B = f\Delta E_{ST} - \Delta E_\pi^\ddagger$. [e] B_X (VB) is derived from eq 6. [f] MUE = 0.6 kcal·mol^{-1}.

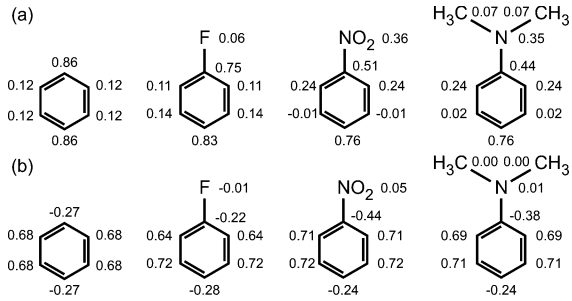


Figure 9. Spin density distributions in the triplet states of a few Ar-X molecules in: (a) the $\pi_y \rightarrow \pi^*_y$ and (b) the $\pi_x \rightarrow \pi^*_x$ states.

density is still smaller than the para value in Figure 9a. It is clear therefore from eqs 4 and 6 that the barriers will be generally larger for meta position attack on the ring, which is what the calculations here and elsewhere[9c,d] generally show, albeit not always.[9d]

The VB barriers of the para attacks (seventh column in Table 2) were modeled using eqs 4 and 6 with experimental IP_X values, $EA_{Cpd\ I} = 64.9$ kcal·mol^{-1} and $f = 0.3$. Thus, for example, using the $IP_H - EA_{Cpd\ I}$ for benzene (148.2 kcal·mol^{-1}), the $IP_X - EA_{Cpd\ I}$ for X = NMe_2 (109.2 kcal·mol^{-1}), and the corresponding ρ_H and ρ_{NMe2} values (0.86 and 0.76, respectively), we obtain $B_{NMe2} = 17.4$ kcal·mol^{-1} and the corresponding barrier 9.4 kcal·mol^{-1} compared with the DFT-calculated datum of 9.6 kcal·mol^{-1}. Other data in Table 2 (columns sixth and seventh) were derived in the same manner.

We note that Table 2 is one of a few almost equally successful modeling sets made with $f = 0.3$ and 1/3 and theoretical and experimental IP_X values; all these attempts give very similar results and are relegated to the Supporting Information (Tables S10−12). As shown by the data in Table 2, the VB barriers (column seventh) model the DFT results (column eighth) quite well with a mean unsigned error of 0.6 kcal·mol^{-1}. Moreover, the trends in the B_X values (sixth column) modeled by eq 6 are close to the values that are required to reproduce the DFT barriers (eq 5). Thus, our modeling of B_X as a quantity based on the mixing of the charge-transfer state into the transition state appears to be quite reasonable and consistent. In both series, the largest B_X is found for X = NMe_2, in good accord with the finding of the largest Q_{CT} for this substituent in Figure 3. Similarly, the smallest B_X is found for X = NO_2, in agreement with the smallest Q_{CT} for this substituent in Figure 3.

The simplest expression to derive the B values for meta attacks is shown in eq 7:

$$B_{X,m} = B_{X,p}[\rho_{X,m}/\rho_{X,p}] \quad (7)$$

which relates the $B_{X,m}$ value to corresponding para value, assuming that the only factor that varies is the relative spin densities in these positions, $\rho_{X,m}/\rho_{X,p}$, all else being constant. Since the meta attack will mix the two triplet states, $\pi_y \rightarrow \pi^*_y$ and $\pi_x \rightarrow \pi^*_x$, we can use the corresponding spin densities in Figure 9b vs a, to derive the $B_{X,m}$ values. Using eq 7, the predicted meta attack barriers are larger than those for the para attack by 1.5−4.9 kcal·mol^{-1}, whereas the corresponding DFT values are 0.4−2.1 kcal·mol^{-1} (except X = NMe_2; see Table 1). In the most deviant case, eq 7 predicts a rise of the meta barrier by 4.7 kcal·mol^{-1} vis-à-vis the DFT calculated 7.2 kcal·mol^{-1}. Obviously eq 7 yields the correct direction in the barrier change, because $B_{X,m} < B_{X,p}$, but it certainly is much oversimplified to provide exact changes in the barrier.

VB Modeling of Reactivity for MeO· Attacks. The VB diagram for arene activation by MeO$^\bullet$ is shown in Figure 10. Here we are concerned only with the π-activation step, which bears similarities to the activation by Cpd I. Indeed, as in Figure 8, here we find the state where the ArX molecule is excited to a triplet and is coupled via a C$^\bullet$−$^\bullet$O bond pair and the charge-transfer state, $^2\Psi_{r,CT}$*, which will mix with the other state curves and generate a TS_π species with a mixed character. However, since MeO$^\bullet$ has a low EA (36.5 kcal·mol^{-1})[20] compared with Cpd I, the charge-transfer state is high lying here in Figure 10 and will mix to a smaller extent into the TS wave function. This is indeed born out by the DFT calculations in Figures 3 vs 4, which shows that the Q_{CT} quantities are always smaller in the MeO$^\bullet$ transition states.

The barrier can be modeled using eq 4 with $f = 0.3$ (see Supporting Information, Tables S14−16). The various B_X values can be derived from eq 4 by using the DFT barriers, and alternatively, it can be modeled using eq 8:

$$B_X = B_H\{[\rho_X(IP_H - EA_{MeO})]/[\rho_H(IP_X - EA_{MeO})]\} \quad (8)$$

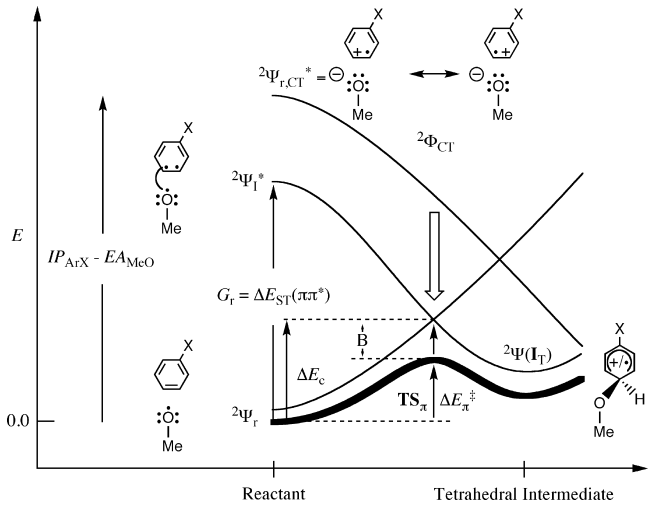


Figure 10. A VB diagram describing the π-activation step for MeO$^\bullet$ attack on the para position of arenes.

Table 3. Reactivity Factors and VB Estimated *B* Values and Barriers for para Position Attacks by MeO$^\bullet$ on X-Substituted Benzene Derivatives[a]

X	B_X[b] (DFT)	B_X[c] (VB)	$\Delta E_\pi^\ddagger$ (VB)[d] (kcal·mol^{-1})	$\Delta E_\pi^\ddagger$ (DFT) (kcal·mol^{-1})
H	22.8	22.8	7.9	7.9
Cl	21.8	21.6	7.5	7.4
F	23.0	21.9	8.7	7.5
CN	19.2	18.6	8.2	7.6
NO_2	18.1	18.2	7.9	8.1
NMe_2	23.5	25.7	1.2	3.3
OMe	23.6	24.3	5.0	5.7
NH_2	23.3	23.8	3.7	4.2
CH_3	22.7	24.7	5.0	7.0
SMe	21.6	23.1	3.8	5.3
N-acetyl	21.3	22.1	5.1	5.9

[a] IP's and $\Delta E_{ST}(\pi\pi^*)$ are those which are given in Table 2. [b] B_X (DFT) is defined by eq 5; $B = f\Delta E_{ST} - \Delta E_\pi^\ddagger$. [c] B_X (VB) is derived from eq 8. [d] MUE = 1.0 kcal·mol^{-1}.

which is analogous to eq 6, with one difference that the EA of Cpd I is replaced by that of MeO$^\bullet$ (calculated vertical EA_{MeO} = 32.1 kcal·mol^{-1}).

All the data are assembled in Table 3, which shows a few trends: First, the B_X values are invariably larger than those for the Cpd I series, in accord with the tighter transition states produced by DFT, in Figure 4 vs 3.

The B_X values estimated from eq 8 and the corresponding VB barriers are close to the DFT-derived ones within a mean unsigned error of 1.0 kcal·mol^{-1}. Furthermore, using eq 7 for the meta B_X values leads to the same conclusion as before, namely that $B_{X,m} < B_{X,p}$. Thus, the VB modeling captures the essence of the two reactions and shows their close relationships.

Addition of Cpd I or MeO$^\bullet$ to aromatic compounds (ArX) generates transition states with similar electronic structures on the substrate, involving both charge and radical characters. Inspection of the transition states (Figures 3 and 4) shows that spin localization is not affected by the reagent identity, whereas the charge-transfer values are more pronounced for transition states generated by Cpd I attack. This larger charge transfer can be attributed to higher EA of Cpd I compared to the methoxy radical. Similar transition-state characters for both reagents are also reflected in the VB modeling. The transition-state energies are gauged by the singlet−triplet energy gap and the resonance energy B_X; the latter quantity reflects also influence by the ability of a substrate to give off an electron and the ability of an oxidant to accept an electron. The VB model thus successfully describes both reaction types and predicts activation energies in reasonably good agreement with the DFT values. It further reveals that the methoxy radical can mimic Cpd I for studies of oxygen addition, as assumed in the pioneering study of Jones et al.[8a]

Table 4. Predicted Activation Energies in kcal·mol^{-1} for the Cpd I Addition to Halogenated Anilines and Benzenes

substrate	position	predicted (VB)[d]	calculated (DFT)
C_6F_6		9.4[a]	8.9[b]
C_6Cl_6		14.8[a]	15.0[b]
1,2-difluoro-benzene	3, 6	16.6	16.6[c]
1,2-difluoro-benzene	4, 5	18.8	15.2[c]
aniline	4	11.7	11.0[c]
2-fluoro-aniline		13.7	11.7[c]
2,6-difluoro-aniline		11.7	12.0[c]
2,3,6-trifluoro-aniline		12.9	11.8[c]

[a] Using Cpd I (HS^-). To gauge the *B* values for these substrates, the VB model used $EA_{Cpd\ I}$ = 67.9 kcal·mol^{-1} for Cpd I (HS^-) and the corresponding barrier for benzene activation (15.8 kcal·mol^{-1}) calculated with UB3LYP/LACV3P(Fe)/6-311+G**(rest)//UB3LYP/LACV3P(Fe)/6-31G*(rest). Using the Cpd I (CH_3S^-) leads to 14.6 and 9.7 kcal·mol^{-1}. [b] See ref 32. [c] See ref 9c. [d] MUE = 1.3 kcal·mol^{-1}.

Making Independent Predictions Using the VB model. The VB model allows us to try making independent predictions of activation energies and compare these with values calculated by DFT in the literature. Some of these predictions are collected in Table 4.

Thus, as can be seen from Table 4, Hackett et al.[32] calculated by DFT, barriers of 15.0 and 8.9 kcal·mol^{-1}, respectively, for oxygenation of C_6Cl_6 and C_6F_6 by Cpd I too while using the singlet−triplet excitation energies and eq 6, the VB model predicts with good agreement, the values of 14.8 kcal·mol^{-1} and 9.4 kcal·mol^{-1} (C_6Cl_6 and C_6F_6). Further, Bathelt et al.[9c] calculated DFT barriers for oxidation of halogenated aniline and 1,2-difluoro-benzene, which compared favorably with experimental data.[5b] It is apparent from Table 4 that the VB model predicts the barriers are in reasonable agreement with DFT.

Whereas the model predicts reasonable barrier values, its ability to predict regioselectivity is somewhat less effective. Thus, in the case of 1,2-difluoro-benzene, Bathelt et al.[9c] predicted a regioselectivity ratio (4,5/3,6) of 63/37 in agreement with the experimental result (67/33).[5b] The VB model predicts on the other hand a preference for 3,6. In other substrates studied theoretically and experimentally,[9d,33] the VB model predicted correctly two cases ((4/6)-methyl-3-fluoro-aniline) and incorrectly for the other two (3-fluoro-aniline and 2-methyl-3-fluoro-aniline). Part of the problem originates in the difficulties to describe all the closely lying triplet states of polysubstituted substrates with DFT. Another part is, of course, due to the fact that regioselectivity is often determined by barrier differences of sub kcal·mol^{-1}, while the accuracy of the VB model is of the order of ~1 kcal·mol^{-1}. One can think of ways to improve the predictive

ability of model by averaging the spin densities for all the closely lying states, but this will affect the simplicity and clarity of the model.

Conclusion

The above study describes a valence-bond modeling approach to the mechanism of arene activation by P450 Cpd I and a methoxy radical. Interestingly, while the VB model is applied here in a manner reminiscent of the quantitative structure−activity relationship methodology,[9e] the VB parameters derive from first principles of electronic structure and, as such, are not arbitrary but rather have physical significance. Thus, VB modeling shows the origins of the barriers for both reaction series and the nature of the corresponding transitions states. Additionally, it elucidates the underling reasons for the stepwise mechanism and the spin selectivity for the reactions of Cpd I. Finally, the model is used to predict barriers for the rate-determining π activation step, with reasonable accuracy compared to the DFT values (mean unsigned deviation for 35 barriers is 1.0 kcal·mol^{-1}), based on easily accessible properties, such as the IPs and singlet−triplet excitation energies of the substrates and the EA of Cpd I. Much order is thus provided by the VB model into P450 chemistry.[11]

Acknowledgment. The research at the Hebrew University is supported by the Israeli Science Foundation (ISF grant 09/53). The authors thank Prof. J. N. Harvey for providing *x*,*y*,*z* coordinates of transition states for Cpd I (CH_3S^-) addition and for kind responses to queries by P.M. P.S. acknowledges the Wenner-Gren Foundation for financial support. Dedicated to Z. Havlas on occasion of his forthcoming 60th birthday.

Supporting Information Available: Cartesian coordinates of the all structures described in this work, Tables with group spin densities and charges, and figures with various correlations are posted. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Sono, M.; Roach, M. P.; Coulter, E. D.; Dawson, J. H. *Chem. Rev.* **1996**, *96*, 2841–2887. (b) *Cytochrome P450: Structure, Mechanism and Biochemistry*, 3rd ed.; Ortiz de Montellano, P. R., Ed.; Kluwer Publishers/Plenum Press: New York, 2005. (c) Guengerich, F. P. *Chem. Res. Toxicol.* **2001**, *14*, 611–650. (d) Ortiz de Montellano, P. R.; De Voss, J. J. *Nat. Prod. Rep.* **2002**, *19*, 477–494. (e) Guengerich, F. P. *Annu. Rev. Pharmacol. Toxicol.* **1999**, *39*, 1–17. (f) Groves, J. T. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3569–3574.

(2) See, e.g.: (a) Groves, J. T. *J. Inorg. Biochem.* **2006**, *100*, 434–447. (b) Schlichting, I.; Berendzen, J.; Chu, K.; Stock, A. M.; Maves, S. A.; Benson, D. E.; Sweet, R. M.; Ringe, D.; Petsko, G. A.; Sligar, S. G. *Science* **2000**, *287*, 1615–1622. (c) Groves, J. In *Cytochrome P450: Structure, Mechanism, and Biochemistry*, 3rd ed.; Ortiz de Montellano, P. R., Ed.; Springer: New York, 2005; Ch. 1, pp 1−43. (d) Denisov, I. G.; Makris, T. M.; Sligar, S. G.; Schlichting, I. *Chem. Rev.* **2005**, *105*, 2253–2277. (e) Ortiz de Montellano, P. R. In *Cytochrome P450: Structure, mechanisms, and biochemistry*, 2nd ed.; Ortiz de Montellano, P. R., Ed.; Plenum Press: New York, 1995; Ch. 8, pp 245−303.

(3) (a) Shaik, S.; de Visser, S. P. In *Cytochrome P450: Structure, Mechanisms, and Biochemistry*, 3rd ed.; Ortiz de Montellano, P. R., Ed.; Springer: New York, 2005; Ch. 2, pp 45−85. (b) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279–2328. (c) Yoshizawa, K. *Coord. Chem. Rev.* **2002**, *226*, 251–259. (d) Meunier, B.; de Visser, S. P.; Shaik, S. *Chem. Rev.* **2004**, *104*, 3947–3980. (e) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 949–1017.

(4) (a) Boyland, E.; Sims, P. *Biochem. J.* **1965**, *95*, 780–787. (b) Jerina, D.; Daly, J.; Witkop, B.; Zaltzman-Nirenberg, P.; Udenfriend, S. *Arch. Biochem. Biophys.* **1968**, *128*, 176–183. (c) Jerina, D.; Daly, J. *Science* **1974**, *185*, 573–582. (d) Jerina, D.; Daly, J.; Witkop, B.; Zaltzman-Nirenberg, P.; Udenfriend, S. *J. Am. Chem. Soc.* **1968**, *90*, 6525–6527. (e) Jerina, D.; Daly, J.; Witkop, B.; Zaltzman-Nirenberg, P.; Udenfriend, S. *Biochemistry* **1970**, *9*, 147–156.

(5) (a) Korzekwa, K. R.; Swinney, D. C.; Trager, W. F. *Biochemistry* **1989**, *28*, 9019–9027. (b) Rietjens, I. M. C. M.; Soffers, A.; Veeger, C.; Vervoort, J. *Biochemistry* **1993**, *32*, 4801–4812. (c) Hanzlik, R.; Hogberg, K.; Judson, C. *Biochemistry* **1984**, *23*, 3048–3055.

(6) (a) Safari, N.; Bahadoran, F.; Hoseinzadeh, M. R.; Ghiasi, R. *J. Porphyrins Phthalocyanines* **2000**, *4*, 285–291. (b) Cnubben, N. H.; Peelen, S.; Borst, J. W.; Vervoort, J.; Veeger, C.; Rietjens, I. M. C. M. *Chem. Res. Toxicol.* **1994**, *7*, 590–598.

(7) (a) Zakharieva, O.; Grodzicki, M.; Trautwein, A. X.; Veeger, C.; Rietjens, I. *J. Biol. Inorg. Chem.* **1996**, *1*, 192–204. (b) Zakharieva, O.; Grodzicki, M.; Trautwein, A.; Veeger, C.; Rietjens, I. *Biophys. Chem.* **1998**, *73*, 189–203.

(8) (a) Jones, J. P.; Mysinger, M.; Korzekwa, K. R. *Drug Metab. Dispos.* **2002**, *30*, 7–12. (b) Dowers, T. S.; Rock, D. A.; Perkins, B. N. S.; Jones, J. P. *Drug Metab. Dispos.* **2004**, *32*, 328–332. (c) Korzekwa, K. R.; Trager, W.; Gouterman, M.; Spangler, D.; Loew, G. H. *J. Am. Chem. Soc.* **1985**, *107*, 4273–4279.

(9) (a) de Visser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2003**, *125*, 7413–7424. (b) Bathelt, C. M.; Ridder, L.; Mulholland, A. J.; Harvey, J. N. *J. Am. Chem. Soc.* **2003**, *125*, 15004–15005. (c) Bathelt, C. M.; Ridder, L.; Mulholland, A. J.; Harvey, J. N. *Org. Biomol. Chem.* **2004**, *2*, 2998–3005. (d) Rydberg, P.; Ryde, U.; Olsen, L. *J. Chem. Theory Comput.* **2008**, *4*, 1369–1377. (e) Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L. *ChemMedChem* **2009**, *4*, 2070–2079.

(10) Bathelt, C. M.; Mulholland, A. J.; Harvey, J. N. *J. Phys. Chem. A* **2008**, *112*, 13149–13156.

(11) (a) Shaik, S.; Kumar, D.; de Visser, S. P. *J. Am. Chem. Soc.* **2008**, *130*, 10128–10140. (b) Shaik, S.; Wang, Y.; Chen, H.; Song, J.; Meir, R. *Faraday Discuss.* **2010**, *145*, 49–70. (c) Shaik, S.; Lai, W.; Chen, H.; Wang, Y. *Acc. Chem. Res.* **2010**, *43*, 1154–1165.

(12) (a) Latifi, R.; Bagherzadeh, M.; de Visser, S. P. *Chem.—Eur. J.* **2009**, *15*, 6651–6662. (b) de Visser, S. P. *J. Am. Chem. Soc.* **2010**, *132*, 1087–1097. (c) de Visser, S. P. *J. Am. Chem. Soc.* **2006**, *128*, 15809–15818. (d) Kumar, D.; Karamzadeh, B.; Narahari Sastry, G.; de Visser, S. P. *J. Am. Chem. Soc.* **2010**, *132*, 7656–7667.

(13) (a) Shaik, S. *J. Am. Chem. Soc.* **1981**, *103*, 3692–3701. (b) Shaik, S.; Shurki, A. *Angew. Chem., Int. Ed.* **1999**, *38*, 586–

625. (c) Shaik, S.; Hiberty, P. C. In *A Chemist's Guide to Valence Bond Theory*; John Wiley & Sons Inc: Hoboken, NJ, 2008; Ch. 6, pp 116−192.

(14) This 'constant' bulk polarity effect can change however, from reaction to reactions, see: de Visser, S. P.; Ogliaro, F.; Sharma, P. K.; Shaik, S. *Angew. Chem., Int. Ed.* **2002**, *41*, 1947–1951.

(15) (a) *Jaguar 4.2* Schrodinger, Inc.: Portland, OR, 2002. (b) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision E.01; Gaussian, Inc.: Wallingford CT, 2004. (c) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; M artin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009. (d) *Jaguar 7.6*; Schrodinger, LLC: New York, 2009.

(16) (a) Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980**, *102*, 7211–7218. (b) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.

(17) (a) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098–3100. (b) Becke, A. D. *J. Chem. Phys.* **1992**, *96*, 2155–2160. (c) Becke, A. D. *J. Chem. Phys.* **1992**, *97*, 9173–9177. (d) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (e) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.

(18) LACV3P is generated in Jaguar 7.6 from LACVP; Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–308.

(19) (a) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724–728. (b) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; DeFrees, D. J.; Pople, J. A.; Gordon, M. S. *J. Chem. Phys.* **1982**, *77*, 3654–3665. (c) Raghavachari, K.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654. (d) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639–5648.

(20) NIST Standard Reference Database No. 69; NIST: Gaithersburg, MD; http://webbook.nist.gov/chemistry/. Accessed March 15, 2010.

(21) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671–6687.

(22) (a) Evans, D. F. *J. Chem. Soc.* **1957**, 3885–3888. (b) Evans, D. F. *J. Chem. Soc.* **1959**, 2753–2757. (c) Yip, R. W.; Sharma, D. K.; Giasson, R.; Gravel, D. *J. Phys. Chem.* **1984**, *88*, 5770–5772. (d) Lim, E. C.; Chakrabarti, S. K. *J. Chem. Phys.* **1967**, *47*, 4726–4730.

(23) (a) Hajgató, B.; Szieberth, D.; Geerlings, P.; De Proft, F.; Deleuze, M. S. *J. Chem. Phys.* **2009**, *131*, 224321. (b) de Silva, E. C.; Gerratt, J.; Cooper, D. L.; Raimondi, M. *J. Chem. Phys.* **1994**, *101*, 3866–3887.

(24) (a) Čížek, J. *J. Chem. Phys.* **1966**, *45*, 4256–4266. (b) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

(25) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(26) (a) Schröder, D.; Shaik, S.; Schwarz, H. *Acc. Chem. Res.* **2000**, *33*, 139–145. (b) Shaik, S.; Hirao, H.; Kumar, D. *Acc. Chem. Res.* **2007**, *40*, 532–542. (c) Shaik, S.; Filatov, M.; Schröder, D.; Schwarz, H. *Chem.—Eur. J.* **1998**, *4*, 193–199.

(27) Carter, E. A.; Goddard III, W. A. *J. Phys. Chem.* **1988**, *92*, 2109–2115.

(28) In $^2\Phi_{COV}$ one electron from the oxo doubly occupied orbital (corresponding to π_{FeO} in Scheme 3a) is shifted to porphyrin$^{+\bullet}$, while the arene in turn is promoted to a triplet state. As such creating $^2\Phi_{COV}$ is quite energy demanding, and the CT state is dominated by the charge-transfer structures.

(29) Such a CT state as in Fig. 7b was tested by the VB method in the following paper, using the two O-C bonds of oxirane: Chen, Z.; Song, J.; Shaik, S.; Hiberty, P. C.; Wu, W. *J. Phys. Chem. A* **2009**, *113*, 11560–11569.

(30) The deformation energy plays a key role in a related energy decomposition analysis scheme. See: (a) Zeist, W.-J.; Bickelhaupt, F. M. *Org. Biomol. Chem.* **2010**, *8*, 3118–3127. (b) Bickelhaupt, F. M. *J. Comput. Chem.* **1999**, *20*, 114–128. (c) de Jong, G. T.; Bickelhaupt, F. M. *ChemPhysChem* **2007**, *8*, 1170–1181.

(31) Su, P.; Song, L.; Wu, W.; Hiberty, P. C.; Shaik, S. *J. Am. Chem. Soc.* **2004**, *126*, 13539–13549.

(32) Hackett, J. C.; Sanan, T. T.; Hadad, C. M. *Biochemistry* **2007**, *46*, 5924–5940.

(33) Koerts, J.; Boeren, S.; Vervoort, J.; Weiss, R.; Veeger, C.; Rietjens, I. M. C. M. *Chem.-Biol. Interact.* **1996**, *99*, 129–146.

CT100554G

JCTC Journal of Chemical Theory and Computation

Kohn−Sham Density Functional Theory Electronic Structure Calculations with Linearly Scaling Computational Time and Memory Usage

Elias Rudberg,*,† Emanuel H. Rubensson,† and Paweł Sałek‡

Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden, and PS Consulting, ul. Zaporoska 8/4, 30-389 Kraków, Poland

Received October 26, 2010

Abstract: We present a complete linear scaling method for hybrid Kohn−Sham density functional theory electronic structure calculations and demonstrate its performance. Particular attention is given to the linear scaling computation of the Kohn−Sham exchange-correlation matrix directly in sparse form within the generalized gradient approximation. The described method makes efficient use of sparse data structures at all times and scales linearly with respect to both computational time and memory usage. Benchmark calculations at the BHandHLYP/3-21G level of theory are presented for polypeptide helix molecules with up to 53 250 atoms. Threshold values for computational approximations were chosen on the basis of their impact on the occupied subspace so that the different parts of the calculations were carried out at balanced levels of accuracy. The largest calculation used 307 204 Gaussian basis functions on a single computer with 72 GB of memory. Benchmarks for three-dimensional water clusters are also included, as well as results using the 6-31G** basis set.

1. Introduction

Recent developments of linear scaling algorithms together with the availability of larger computer resources have made it possible to carry out electronic structure calculations for systems with many thousands of atoms,[1−6] using Hartree−Fock (HF) and Kohn−Sham density functional theory (KS-DFT), as well as tight-binding and MNDO-type semiempirical methods. Linear scaling algorithms have been developed for all computationally expensive parts of such calculations, including the computation of the Coulomb,[7−9] HF exchange,[10,11] and KS-DFT exchange-correlation[12,13] matrices and methods for the density matrix construction step.[14−18] However, making efficient use of linear scaling methods is not straightforward, and quadratically or worse scaling methods are still applied even for molecular systems with many thousands of atoms, where the use of linear scaling algorithms would greatly reduce the computational cost, see, e.g., ref 19.

The development of linear scaling algorithms has mostly focused on achieving linear scaling in time, and the memory usage aspect has often been overlooked. Whereas linear scaling in memory can in principle be achieved by replacing dense matrix storage with a sparse matrix representation, the overhead from locating and modifying elements in the sparse storage can be considerable. Often, linear scaling algorithms have been described without considering the matrix representation, thus implicitly assuming fast access to matrix elements. In practice, efficient use of sparse matrix storage requires that the way that matrix elements are accessed is considered as an integral part of each algorithm and that the different parts of the calculation are combined efficiently while using only sparse data structures. Also, in order to achieve sufficient accuracy in the result, the various threshold values used in different parts of the calculation need to be chosen carefully. Due to such difficulties, linear scaling methods are still far from reaching their full potential.

* Corresponding author e-mail: elias.rudberg@it.uu.se.
† Uppsala University.
‡ PS Consulting.

10.1021/ct100611z
Published on Web 12/30/2010

This paper builds on our earlier work on linearly scaling HF calculations.[4] We describe the necessary changes needed to efficiently evaluate the exchange-correlation contributions to the Kohn−Sham matrix with linear scaling in both time and memory. In particular, an efficient way of accessing sparse matrix elements is described, with an overhead comparable to that of dense matrix storage. The new algorithm has been implemented in the Ergo program.[20] We also describe how the exchange-correlation contributions can be evaluated at an accuracy level consistent with the other parts of the calculation.

This paper is organized as follows: Section 2 gives an overview of the KS-DFT method from a computational point of view. Our algorithm for linear scaling construction of the exchange-correlation matrix directly in sparse form is described in section 3. Benchmark calculations demonstrating linear scaling behavior are presented in section 4. Finally, a few concluding remarks are given in section 5.

2. Method

We consider calculations where the electron density is expanded in a set of n basis functions $\{\phi_k(r)\}$ built up by combinations of polynomials and Gaussian functions usually centered at the nuclei of the molecule. These basis sets are commonly referred to as Gaussian type linear combination of atomic orbital (GT-LCAO) basis sets or simply Gaussian basis sets. See ref 21 for a thorough discussion about such basis sets.

The sequence of steps illustrated in Algorithm 1 summarizes how a linear scaling KS-DFT calculation is carried out in the Ergo quantum chemistry program.[20] The Ergo program uses Gaussian basis sets to compute electronic structures with linearly scaling processor time and memory usage.

Algorithm 1.	Overview of KS-DFT Self-Consistent Field Program
1:	Read molecule and basis set information from input files.
2:	Compute overlap matrix S and inverse factor Z such that $Z^TSZ = I$.
3:	Compute one-electron Hamiltonian matrix H_1.
4:	Generate starting guess density matrix D.
5:	**for** $i = 1, 2, \ldots$ **do**
6:	Compute new Coulomb matrix J.
7:	Compute new HF exchange matrix K.
8:	Compute new Kohn−Sham exchange-correlation matrix V_{xc} and energy E_{xc}
9:	Compute energy $E = Tr(DH_1) + {}^1/_2 Tr(D(J + \gamma K)) + E_{xc}$.
10:	Compute new Kohn−Sham matrix $F = H_1 + J + \gamma K + V_{xc}$.
11:	Compute $\tilde{F}$ as a linear combination of new and previous Kohn−Sham matrices.
12:	Compute $F_\perp = Z^T\tilde{F}Z$.
13:	Compute new density matrix $D_\perp$ from $F_\perp$.
14:	Compute $D = ZD_\perp Z^T$.
15:	**end for**

In the self-consistent field (SCF) procedure given by Algorithm 1, two main operations are repeated: (1) the $D \rightarrow F$ step, for the construction of the Kohn−Sham matrix for a given density matrix, consisting of steps 6−10 of Algorithm 1, and (2) the $F \rightarrow D$ step, for the subsequent construction of a new density matrix, consisting of steps 12−14. These two operations can be employed in a simple fixed point iteration, but usually some convergence enhancing schemes are used to accelerate and hopefully even ensure convergence, see refs 22 and 23 for recent reviews. In our calculations, in each iteration, either damping[24,25] or DIIS[26,27] is used in step 11 to generate $\tilde{F}$ as a linear combination of new and previous Kohn−Sham matrices.

The Kohn−Sham matrix F consists of one-electron (H_1) and two-electron (J, K, V_{xc}) contributions: $F = H_1 + J + \gamma K + V_{xc}$. In the case of a HF calculation, $V_{xc} = 0$, $E_{xc} = 0$, and $\gamma = 1$. In the case of a pure Kohn−Sham calculation, $\gamma = 0$. For so-called hybrid functionals, V_{xc} and E_{xc} are both nonzero and $\gamma \neq 0$. Becke's half-and-half functional with an LYP correlation part (BHandHLYP),[28] used in the benchmark calculations described in section 4, is a hybrid functional with $\gamma = 0.5$.

The Coulomb matrix J can be efficiently calculated using truncated multipole expansions.[8,9,29−32] An important feature of our implementation is the use of a dynamically selected multipole expansion order, an approach that gives significant speedups compared to always using the same expansion order.[8,9] Truncated multipole expansions are also used for linear scaling computation of the electron-nuclei term of the one-electron Hamiltonian matrix H_1.

The HF exchange matrix K can be computed in linear time by exploiting the locality of basis functions together with the sparsity of the density matrix. There has been much research devoted to efficient computation of the exchange matrix.[10,11,33−38] Some details about the exchange matrix evaluation in the Ergo code, including memory usage considerations, can be found in ref 4.

A key result of this article is our algorithm for linear scaling construction of the Kohn−Sham exchange-correlation matrix V_{xc} directly in sparse form. This algorithm is described in section 3.

The density matrix $D_\perp$ is computed from $F_\perp$ using the purification scheme of ref 18 combined with a novel approach for the removal of small matrix elements.[39] The most distinguished feature of this purification scheme, which uses the so-called trace-correcting purification polynomials of the second order,[17] is that it allows for rigorous control of the error in the occupied subspace. This purification procedure is formulated for an orthogonal basis. Therefore, an inverse factor Z of the overlap matrix is needed for the congruence transformations to and from the orthogonal basis in steps 12 and 14 of Algorithm 1. In the calculations considered in this paper, the inverse factor was computed using inverse Cholesky decomposition,[40−42] although other choices for Z are possible as well.[43−45]

Matrix operations needed particularly in density matrix purification but also for other operations are performed using sparse matrix algebra. This allows for linear scaling provided that the matrix sparsity is such that the average number of nonzero elements per row does not increase with system size. This is usually the case for large nonmetallic molecular systems.

For simplicity, the description above was given for the common case of a spin-restricted calculation. However, this can be straightforwardly generalized to the spin-unrestricted

case. Then, the electron densities for α and β spin are represented by separate density matrices D_α and D_β so that the total electron density matrix is given by $D = D_\alpha + D_\beta$, and similarly two Kohn−Sham matrices F_α and F_β are created. Two Kohn−Sham exchange-correlation matrices $V_{xc;\alpha}$ and $V_{xc;\beta}$ are also used, formed from the electron densities ρ_α and ρ_β as described in the following section.

3. Linear Scaling Computation of the Kohn−Sham Exchange-Correlation Matrix

The Kohn−Sham formulation of density functional theory[46] allows one to formulate the framework for density functional theory calculations in a way similar to the HF framework with two important modifications. HF exchange is scaled down or entirely removed. Instead, an exchange-correlation term is added to the energy, and a corresponding contribution is added to the Fock matrix. A Fock matrix with an exchange-correlation contribution is traditionally called a Kohn−Sham matrix.

The exchange-correlation energy E_{xc} within the generalized gradient approximation (GGA) is given by

$$E_{xc} = \int_{\mathbb{R}^3} \mathscr{F}(\rho_\alpha(r), \rho_\beta(r), q_\alpha(r), q_\beta(r), q_{\alpha\beta}(r))\, dr \quad (1)$$

where

$$q_\alpha(r) = |g_\alpha(r)| \quad (2)$$

$$q_{\alpha\beta}(r) = g_\alpha(r) \cdot g_\beta(r) \quad (3)$$

$$g_\alpha(r) = \nabla \rho_\alpha(r) \quad (4)$$

Here, $\mathscr{F}(\rho_\alpha(r),\rho_\beta(r),q_\alpha(r),q_\beta(r),q_{\alpha\beta}(r)) \equiv \mathscr{F}(r)$ is the density functional which in the case of GGA also depends on the density gradient.

GGA assumes that the nonlocal character of the exchange-correlation contributions can be captured by making the functional dependent on the local value of the spin-dependent density gradient $g_\alpha(r)$. It also separately considers electron densities with spin-up α and -down β, similarly as in the local spin density approximation. The densities ρ_α and ρ_β are the same only in the special case of a closed shell calculation.

Matrix elements of the exchange-correlation matrix $V_{xc;\alpha}$ under the GGA approximation are given by

$$V_{xc;\alpha;pq} = \int_{\mathbb{R}^3} [s_{\alpha;p}(r)\, \phi_q(r) + \phi_p(r)\, s_{\alpha;q}(r)]\, dr \quad (5)$$

where

$$s_{\alpha;p}(r) = \phi_p(r)\, v_\alpha(r) + u_\alpha(r) \sum_{c \in \{x,y,z\}} \frac{\partial \phi_p(r)}{\partial c} \frac{\partial \rho_\alpha}{\partial c} + t(r) \sum_{c \in \{x,y,z\}} \frac{\partial \phi_p(r)}{\partial c} \frac{\partial \rho_\alpha}{\partial c} \quad (6)$$

$$v_\alpha(r) = \frac{\partial \mathscr{F}(r)}{\partial \rho_\alpha} \quad (7)$$

$$u_\alpha(r) = \frac{1}{q_\alpha(r)} \frac{\partial \mathscr{F}(r)}{\partial q_\alpha} \quad (8)$$

$$t(r) = \frac{\partial \mathscr{F}(r)}{\partial q_{\alpha\beta}} \quad (9)$$

In contrast to the integrals encountered in calculations of Coulomb repulsion and HF exchange, exchange-correlation integrals cannot be evaluated using a compact analytical expression. Instead, the exchange-correlation energy E_{xc} and the matrix elements $V_{xc;\alpha;pq}$ are computed using numerical integration over a grid:

$$E_{xc} = \sum_i w_i \mathscr{F}(r_i) \quad (10)$$

$$V_{xc;\alpha;pq} = \sum_i w_i [s_{\alpha;p}(r_i)\, \phi_q(r_i) + \phi_p(r_i)\, s_{\alpha;q}(r_i)] \quad (11)$$

The choice of grid point locations $\{r_i\}$ and associated grid weights $\{w_i\}$ determines the quality of the integration grid. The electron density at a given grid point r_i is computed by contracting the density matrix D_α with basis functions evaluated at r_i:

$$\rho_\alpha(r_i) = \sum_{pq} D_{\alpha;pq} \phi_p(r_i)\, \phi_q(r_i) \quad (12)$$

The calculation of the exchange-correlation matrix formally scales cubically with system size. The scaling can be reduced to linear if basis function screening is implemented.

3.1. Grids for Numerical Integration. Traditionally, the entire integration grid is constructed as a union of atomic grids, with grid weights $\{w_i\}$ adjusted in the overlapping regions.[47] Atomic grids are constructed as outer products of Lebedev grids for angular integration[48] and Gauss−Chebychev radial grids.[47,49] Alternative radial grids have been proposed as well.[50] The weights in overlapping regions are adjusted using Becke partitioning or its variants.[12,51] Smooth switching functions used in the Becke partitioning process in principle stretch out infinitely. This makes the computational cost of the partitioning process scale cubically with system size. In practice, the right choice of multiplication order used in the grid weight scaling process can make the effort per atom roughly independent of the system size. Other partitioning schemes, see, e.g., ref 12, choose the partitioning function in a way that allows for trivial screening of atoms far away from the grid point associated with the weight being adjusted.

While grids constructed as unions of atomic grids are well established, the existence of overlapping regions in multiatom systems introduces errors that are difficult to control. The high accuracy that is possible for the integration of densities or exchange-correlation potentials for spherically symmetric systems like atoms cannot be realized in such cases. The complication of overlapping regions makes the error increase by several orders of magnitude. A grid construction method that in principle allows for integration of the electron density

up to any accuracy is the so-called hierarchical cubature (HiCu).[13] In the HiCu scheme, the quality of the local integration grid in each part of space is measured by comparing the numerical integral of the electron density to the corresponding analytically evaluated integral. The grid is further refined until a predefined integration accuracy criterion is met. Other parameters such as basis function extents are related to the integration accuracy criterion, resulting in a method where a single parameter is used to control the accuracy of the computed exchange-correlation matrix.

In the Ergo program, both space partitioned atomic grids as well as the HiCu scheme are implemented. In the benchmark calculations in section 4, we used the HiCu scheme since it has the advantage that the accuracy is controlled by a single parameter.

3.2. Evaluation of the Sparse Exchange-Correlation Potential Matrix. The evaluation of the exchange-correlation matrix as given by eq 5 formally follows the scheme shown in Algorithm 2. An efficient implementation of that algorithm must fulfill a few conditions:

(1) Matrix element magnitudes $|V_{xc;\alpha;pq}|$ are estimated in advance so that memory for $V_{xc;\alpha;pq} < \tau$ is not allocated and the sum contributing to that element is not computed. Here, τ is a preselected threshold for matrix elements.

(2) Terms $s_{\alpha;q}(r_i)$, $\phi_p(r_i)$, and $\phi_q(r_i)$, contributing to several matrix elements, are not unnecessarily recomputed.

Algorithm 2. Numerical Integration
1: **for** each (p,q) giving rise to a nonvanishing $V_{xc;\alpha;pq}$ **do**
2: $V_{xc;\alpha;pq} := \sum_i w_i[s_{\alpha;p}(r_i)\ \phi_q(r_i) + \phi_p(r_i)\ s_{\alpha;q}(r_i)]$
3: **end for**

Depending on the amount of available memory and other considerations, the operations in Algorithm 2 may be performed in a different order. If memory constraints were not present, we could perform the operations as shown in Algorithm 3. This simple algorithm has significant memory requirements. The sparse matrix B needs to be available during the entire integration process. Let us consider for example a system with 10 000 atoms, with 10 000 grid points per atom, and where on average 50 basis functions are nonvanishing at a grid point. In that example, the matrix B would require approximately 60 GB of memory if stored in the compressed sparse row format.[52] One way to reduce the memory demand is to process the grid points in batches. The Ergo implementation follows that approach.

Algorithm 3. Linearly Scaling Numerical Integration
1: Compute a sparse matrix B with elements $B_{ki} = \phi_k(r_i)$ of basis functions ϕ_k evaluated at grid points r_i.
2: Compute $\rho_\sigma(r_i)$, $\sigma \in \{\alpha, \beta\}$ by contracting the sparse density matrix D_σ with sparse B on each side: $\rho_\sigma(r_i) = \sum_{pq} B_{pi}D_{\sigma;pq}B_{qi}$.
3: Use $\rho_\sigma(r_i)$ to compute $v_\sigma(r_i)$, $u_\sigma(r_i)$, $t(r_i)$ and eventually $S_{\sigma;pi} = s_{\sigma;p}(r_i)$, and store the result.
4: Compute the exchange-correlation matrix by performing a matrix scaling and a sparse matrix-matrix multiplication as given by eq 11: $V_{xc;\sigma;pq} = \sum_i w_i(B_{qi}S_{\sigma;pi} + B_{pi}S_{\sigma;qi})$

At the time of grid generation, grid points are collected in spatial cells. For each cell, we find the basis functions that overlap with that cell. This information is important with respect to both accuracy and performance. A too cautious estimation may result in a dramatic increase in calculation time. On the other hand, a too sloppy criterion for determination of basis functions relevant for a given cell will inadvertently affect the calculation accuracy. The list of nonvanishing basis functions for a given cell allows us to predict, using eq 11, which exchange-correlation matrix elements may have nonzero values and to determine an exchange-correlation matrix sparsity pattern. This pattern in turn permits preallocation of the resulting exchange-correlation matrix $V_{xc;pq}$ so that individual contributions computed later can be added quickly without a need to reallocate memory. The numerical integration is then performed one cell at a time. For each cell, we follow Algorithm 3, with the exception that the partial contributions computed according to this algorithm are accumulated for all cells.

We store the sparse matrix pattern for the exchange-correlation calculation as a list of nonvanishing matrix element ranges for each column of the matrix. For any given column, we choose to number the atoms and associated basis functions following their spatial location, so that the number of ranges is small. This data structure has several properties important from the performance point of view. First, ranges can be rapidly extended as the sparse matrix pattern is built during grid generation. Second, a small number of ranges for any given column, together with an efficient bisection algorithm, makes it possible to efficiently find elements in the sparse matrix. To assess the efficiency of our sparse matrix structure, we have compared the performance to a previous version of the code using full dense matrix storage. The overhead of finding matrix elements in the sparse matrix structure does not exceed 10% of the total exchange-correlation integration time in pessimistic cases, and in many cases, the sparse code is faster by 10% to 20% than the corresponding version operating on full matrices due to increased computer memory locality and improved cache usage.

4. Benchmark Calculations

In this section, we present benchmark calculations for two kinds of molecular systems: glutamic acid−alanine (Glu−Ala) helices and water clusters. For each of them, we have performed KS-DFT calculations for varying system sizes using the BHandHLYP functional[28] with two different basis sets: 3-21G and 6-31G**.

4.1. Molecular Systems Used for Benchmarks. The Glu−Ala helix systems were generated using the "build sequence" function in the Spartan program,[53] with the α helix option selected. We refer to these systems as $[GluAla]_n$, where n is the number of repeating Glu−Ala units. Because generating very large helices using the Spartan program became cumbersome, systems larger than $[GluAla]_{448}$ were instead generated from the smaller ones using an elongation procedure.

The water cluster geometries were generated from a large molecular dynamics simulation of bulk water at standard temperature and pressure by including all water molecules within spheres of varying radii. A water molecule was

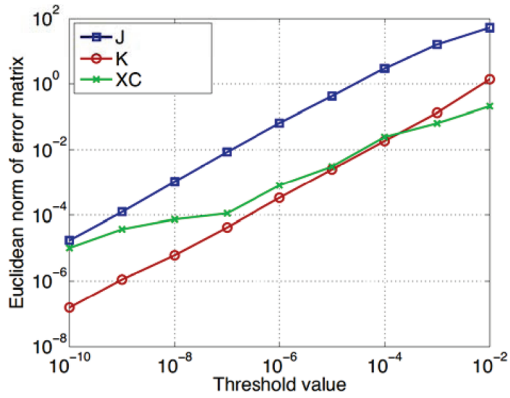


Figure 1. Accuracy scans for BHandHLYP/3-21G calculations on a water cluster containing 471 water molecules, near SCF convergence. These scans are used together with inequality 13 to select threshold values for Coulomb matrix (*J*), HF exchange matrix (*K*), and exchange-correlation matrix (XC) evaluations that correspond to a requested level of accuracy in the occupied subspace.

included if its oxygen atom was within the radius, thus making sure that only whole water molecules were included.

The Glu−Ala and water cluster systems used in the benchmarks are comparable to the ones used in ref 4, although the water cluster geometries were now generated from a different, larger molecular dynamics simulation. Thanks to improvements of our implementation as well as the availability of greater computer memory, we are now able to carry out benchmark calculations for significantly larger systems.

To facilitate comparison and verification of results, all molecular geometries used in this paper are available for download at www.ergoscf.org.

4.2. Selection of Threshold Values. The key quantity of interest in electronic structure calculations is the occupied subspace. Computational approximations result in perturbations of this subspace. Errors can be measured by the largest canonical angle θ_1 between exact and perturbed subspaces. This angle is related to the Euclidean norm of the error matrix E and the gap ξ between eigenvalues corresponding to occupied and unoccupied orbitals respectively:[54]

$$\sin\theta_1 \leq \frac{\|E\|_2}{\xi - \|E\|_2} \tag{13}$$

Note that in the case of a Kohn−Sham matrix, ξ is equal to the HOMO−LUMO gap, and in case of a density matrix, $\xi = 1$. For the benchmark calculations, we selected threshold values such that the erroneous rotation of the occupied subspace in each of the $F\rightarrow D$ and $D\rightarrow F$ steps as measured by $\sin\theta_1$ would be below 1×10^{-2}.

For the $F\rightarrow D$ step, our implementation of density matrix purification allows us to specify the desired accuracy directly in terms of $\sin\theta_1$.[18] For the $D\rightarrow F$ step, we use the accuracy scan information in Figure 1 together with inequality 13 to choose threshold values that correspond to a given error in the occupied subspace. The needed information about the HOMO−LUMO gap is obtained as a byproduct of density matrix purification.[18,55] The resulting threshold values, chosen so that $\sin\theta_1 \leq 1 \times 10^{-2}$ in each of the $F\rightarrow D$ and $D\rightarrow F$ steps for BHandHLYP water cluster calculations, are shown in Table 1. Note that the threshold values τ_J and τ_K, for Coulomb and HF exchange matrix construction, respectively, differ by more than 2 orders of magnitude. This is in spite of the fact that these threshold values are used in a similar way, namely, that contributions smaller than the specified value are neglected. Thus, there is no universal relationship between neglect threshold and error matrix norm.

Table 1. Threshold Values Selected to Give an Accuracy in the Occupied Subspace of $\sin\theta_1 \leq 1 \times 10^{-2}$ in Each of the $F\rightarrow D$ and $D\rightarrow F$ Steps, Where the $D\rightarrow F$ Step Consists of *J*, *K*, and V_{xc} Matrix Evaluations[a]

calculation	threshold value
$F\rightarrow D$ step	$\tau_D = 1 \times 10^{-2}$
Coulomb matrix *J*	$\tau_J = 5 \times 10^{-9}$
HF exchange matrix *K*	$\tau_K = 2 \times 10^{-6}$
exchange-correlation matrix V_{xc}	$\tau_{xc} = 5 \times 10^{-7}$

[a] The values were determined using the accuracy scans in Figure 1 together with eq 13. The HOMO−LUMO gap in the BHandHLYP/3-21G calculation was 0.23 au.

In principle, separate accuracy scans should be performed also for 6-31G** and for the Glu−Ala systems, giving different sets of threshold values for each case. Whereas the different gap value for Glu−Ala is automatically taken into account by the program in the $F\rightarrow D$ step, this is not the case in the construction of the Kohn−Sham matrix. Ideally, the different gap value for Glu−Ala should have been taken into account in the calculations of *J*, *K*, and V_{xc} as well. However, for simplicity, we have used the set of threshold values in Table 1 for all benchmark calculations reported in this work, with one exception: the parameter τ_D determining the accuracy of the $F\rightarrow D$ step was set to 1×10^{-3} in the Glu−Ala calculations, in order to get more accurate information about HOMO and LUMO eigenvalues. As will be seen below, this parameter change for Glu−Ala did not significantly affect the overall computational time since the Glu−Ala calculations were strongly dominated by the $D\rightarrow F$ step.

As can be seen in Figure 1, the errors in *J* and *K* decrease in a very predictable manner as the integral threshold values are decreased. It is therefore possible to implement a program that automatically selects appropriate threshold values by extrapolation after an assessment of the slopes of the lines.[56] We believe that such an automated procedure in principle should be possible also for the exchange-correlation part. However, with our current implementation, the error in the exchange-correlation part decreases in a less predictable manner, see Figure 1, which makes it difficult to automate the procedure of selecting a threshold value for the exchange-correlation matrix evaluation.

4.3. Results. Results of the benchmark calculations are shown in Figures 2, 3, 4, and 5. Each figure includes timings, matrix sparsity, memory usage, and the HOMO−LUMO gap plotted against system size. The calculations were performed using the Ergo program,[20] compiled with the Intel C++ compiler (ICC), version 10.1, and linked to the GotoBLAS2 linear algebra library, version 1.11p1.[57,58] Each calculation was run on a HP SL170h G6 compute server with two Quad-

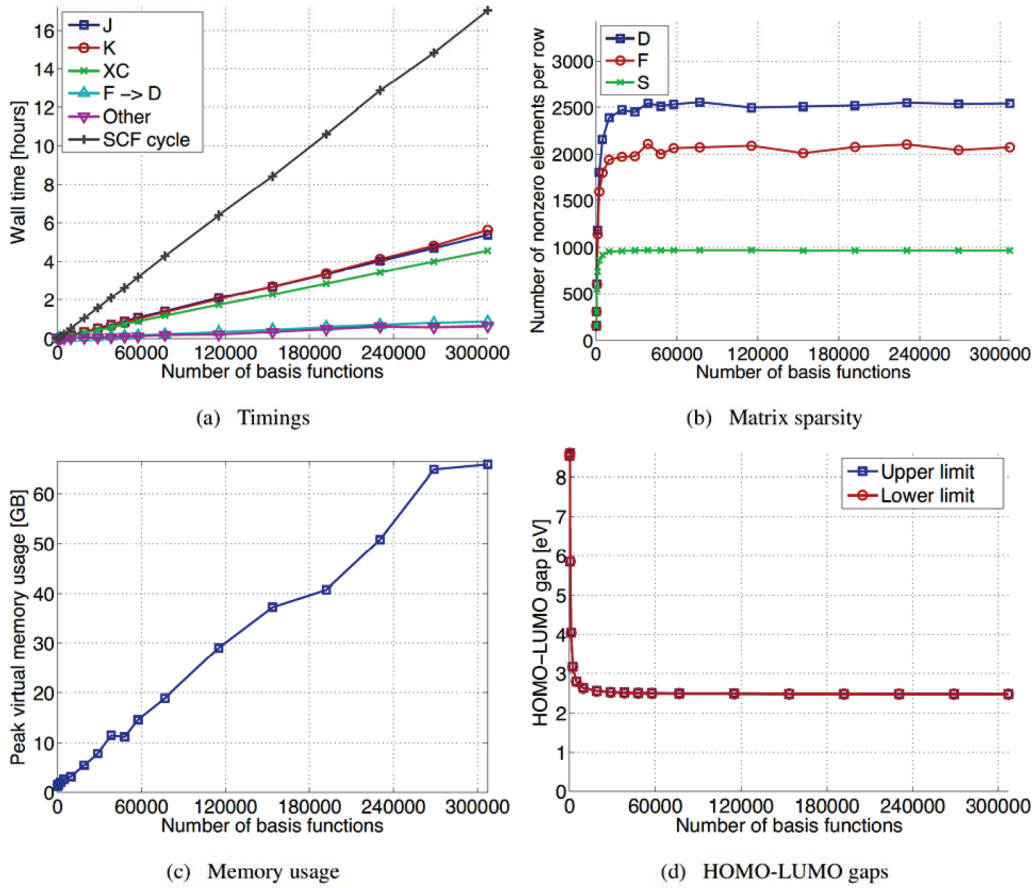


Figure 2. Timings, memory usage, matrix sparsity, and computed HOMO−LUMO gaps for BHandHLYP/3-21G calculations on Glu−Ala helix systems of varying size. The right-most points in the graphs are for $[GluAla]_{2048}$, with 53 250 atoms and 307 204 basis functions. The upper and lower limits of the HOMO−LUMO gaps are indistinguishable in the figure.

core Intel Xeon 5520 (Nehalem 2.26 GHz, 8 MB cache) processors and 72 GB of shared memory running the Scientific Linux 5.4 operating system. Block-sparse matrix operations were performed using a version of the Hierarchic Matrix Library[42,59] parallelized for shared memory using OpenMP, with a uniform block size of 32 at the lowest level. Other time-critical parts of the code are parallelized using POSIX threads.

To generate starting guess densities, we first performed preliminary calculations at the HF/STO-2G level of theory. The resulting densities were used as starting guesses for calculations at the HF/3-21G level. Finally, the converged HF/3-21G densities were used as starting guesses for the BHandHLYP calculations for both the 3-21G and 6-31G** basis sets. The HF/3-21G guesses were good enough for the BHandHLYP calculations to converge within 4−6 self-consistent field iterations. The calculations were considered converged as soon as the largest magnitude element of $FDS - SDF$ was smaller than 1×10^{-3}.

The plotted timings do not include the initialization work that is needed before the first SCF cycle, i.e., steps 1−5 in Algorithm 1. Those steps contributed only to a small part of the total calculation time. The whole initialization work, including computation of the overlap and one-electron Hamiltonian matrices, inverse Cholesky decomposition, and starting guess density projection parts, in no case took more than 12% of the total calculation time.

The grid for numerical integration in the KS-DFT exchange-correlation matrix evaluation was created using the HiCu method[13] with threshold value $\tau_{xc} = 5 \times 10^{-7}$, which gave on average around 9100 grid points per atom for the Glu−Ala calculations and around 7100 grid points per atom for the water cluster calculations. The choice of basis set did not significantly affect the number of grid points. Even though roughly the same number of grid points were generated for both basis sets, the grid generation for 6-31G** required about 4−5 times as long time as for 3-21G. This is because the description of the electron density used during the HiCu grid generation[13] becomes more expensive when using a larger basis set. The grid was generated only once, in the first SCF cycle. In subsequent cycles, the same grid was reused. Therefore, the exchange-correlation matrix evaluation in the first SCF cycle was computationally more expensive by roughly a factor of 2. Linear scaling was observed for the grid generation, although this extra time is not seen in the figures as the plotted timings are for the third SCF cycle.

The results of the Glu−Ala helix benchmark calculations, shown in Figures 2 and 3 for basis sets 3-21G and 6-31G**, respectively, indicate nearly perfect linear scaling. The timings are strongly dominated by the $D \rightarrow F$ step, consisting of the J, K, and XC parts, while the $F \rightarrow D$ step requires less than 10% of the total SCF cycle time. For the larger helices, the number of nonzero elements in the density matrix

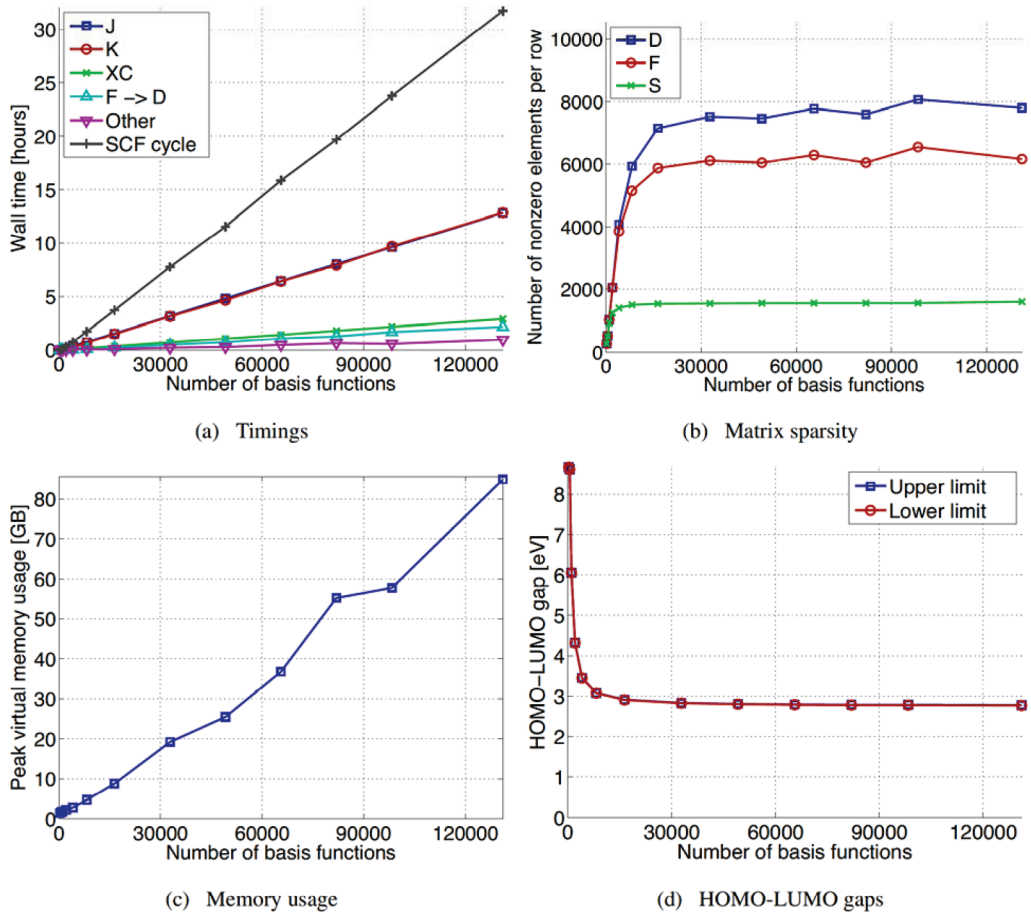


Figure 3. Timings, memory usage, matrix sparsity, and computed HOMO−LUMO gaps for BHandHLYP/6-31G** calculations on Glu−Ala helix systems of varying size. The right-most points in the graphs are for $[GluAla]_{512}$, with 13 314 atoms and 131 082 basis functions. The timings for the Coulomb (*J*) and HF exchange (*K*) parts in panel a are almost identical.

stabilizes at around 2500 and 8000 elements per row for basis sets 3-21G and 6-31G**, respectively. For the 3-21G basis set, the largest helix we could handle was $[GluAla]_{2048}$, $C_{16384}N_{4096}O_{8192}H_{24578}$, corresponding to 307 204 basis functions. For the 6-31G** basis set, the largest helix we could handle was $[GluAla]_{512}$, $C_{4096}N_{1024}O_{2048}H_{6146}$, corresponding to 131 082 basis functions.

In the water cluster benchmark calculations, shown in Figures 4 and 5 for basis sets 3-21G and 6-31G**, respectively, linear scaling behavior is approached for the larger systems. Perfect linear scaling is expected when reaching system sizes where the number of nonzero elements per row in the density matrix no longer increases. We note that for the larger water cluster calculations the $F \rightarrow D$ step takes a considerable part of the total SCF cycle time, and it is clearly the last part of the calculation to enter the linear scaling regime. This is in sharp contrast to the one-dimensional Glu−Ala case where the $F \rightarrow D$ step takes only a small fraction of the total SCF cycle time. For the larger water clusters, the number of nonzero elements in the density matrix approaches 6000 and 13 000 elements per row for basis sets 3-21G and 6-31G**, respectively. For the 3-21G basis set, the largest water cluster we could handle contained 9644 water molecules, corresponding to 125 372 basis functions. For the 6-31G** basis set, the largest water cluster we could handle contained 3050 water molecules, corresponding to 73 200 basis functions.

The memory usage plotted in panel c of Figures 2, 3, 4, and 5 is the peak virtual memory usage as reported by the operating system. In some of the largest calculations, the virtual memory usage was slightly above the physical memory limit of 72 GB, so that some swapping to disk by the operating system must have occurred. However, this incurred no noticeable overhead.

For both water clusters and Glu−Ala helix systems, the increase in computational cost when changing basis set from 3-21G to 6-31G** was around a factor of 5−7 in computational time and around a factor of 3−4 in memory requirement.

HOMO and LUMO eigenvalues were computed by applying the Lanczos method in intermediate purification iterations as described in ref 18. The resulting HOMO−LUMO gaps are plotted in panel d of Figures 2, 3, 4, and 5. For the larger GluAla systems, the computed HOMO−LUMO gaps are 2.5 and 2.8 eV for 3-21G and 6-31G**, respectively. For the larger water cluster systems, the computed HOMO−LUMO gaps are 4.2−4.9 eV and 5.6−6.4 eV for 3-21G and 6-31G**, respectively. It should be noted that the HOMO−LUMO gaps resulting from this kind of calculation are strongly dependent on the amount of HF exchange in the KS-DFT functional. We chose to use the BHandHLYP functional for these benchmarks in order to get sufficiently large HOMO−LUMO gaps to allow efficient calculations. In fact, for the larger Glu−Ala and

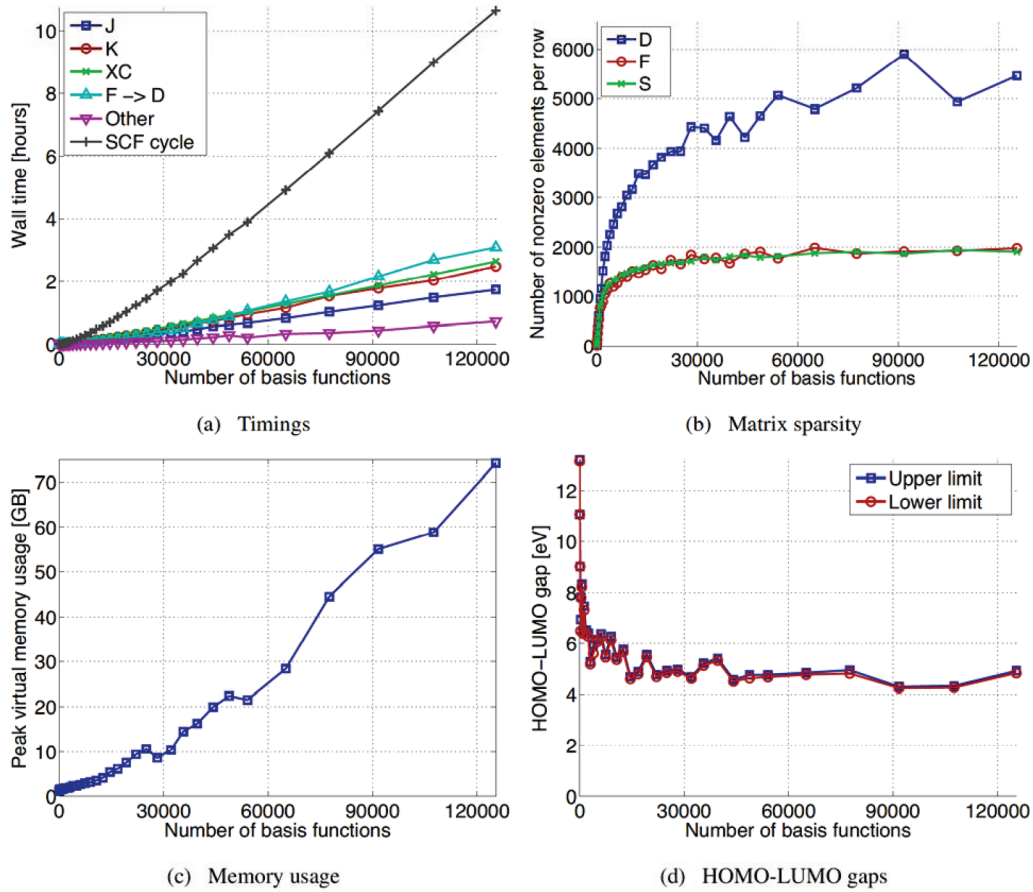


Figure 4. Timings, memory usage, matrix sparsity, and computed HOMO−LUMO gaps for BHandHLYP/3-21G calculations on water cluster systems of varying size. The right-most points in the graphs are for a water cluster containing 9644 water molecules, corresponding to 125 372 basis functions.

water cluster systems as well as for large protein molecules in general, our experience is that some fraction of HF exchange in the functional is necessary in order to get a nonvanishing HOMO−LUMO gap. We have been unable reach convergence in attempted calculations using pure functionals such as LDA, BLYP, and PBE for all but the smallest systems.

In the matrix sparsity plots, the curves are somewhat jagged, particularly in Figure 4b. This is an effect of the stepping procedure in the applied truncation scheme which ensures that the norm of the error matrix is below a requested value, see refs 39 and 60. An adjustment of a stepping parameter in the truncation scheme implementation would make the curves more smooth. The effect is more pronounced in the water cluster calculations because, compared to the Glu−Ala case, more matrix elements are removed in each truncation.

By comparison to higher accuracy results, we have found that, for both water clusters and Glu−Ala, the errors in total energies in the larger calculations were below 1×10^{-5} Hartree/atom for the 3-21G basis set, and below 3×10^{-5} Hartree/atom for the 6-31G** basis set.

5. Concluding Remarks

It should be noted that since we employed a hybrid KS-DFT functional, the calculations presented here include all components needed for HF calculations. Thus, although the presented benchmarks do not include HF calculations, we are able to perform HF calculations for these systems as well. Compared to a BHandHLYP calculation with the same basis set, a HF calculation is typically somewhat faster since the exchange-correlation matrix evaluation is not needed. Also, HF calculations are generally easier to carry out because HF gives a larger HOMO−LUMO gap than DFT.

An important aspect of linear scaling calculations is the selection of threshold values. In the calculations reported in this work, threshold values were chosen to give roughly the same accuracy for the different parts. Threshold values for the different parts of the $D \rightarrow F$ step were chosen using information from previous accuracy scans. Threshold values in the $F \rightarrow D$ step were automatically selected by the program to achieve the requested accuracy. Ideally, threshold values for the $D \rightarrow F$ step should also be chosen automatically, for example using extrapolation.[56] In any case, balancing the accuracy can considerably improve the overall performance. One example of this is the selection of integral screening threshold values for Coulomb and HF exchange matrix evaluation. In our previous study of linear scaling HF calculations,[4] using ad hoc selected threshold values, the HF exchange matrix evaluation was about twice as expensive as the Coulomb part. Now, having adapted the threshold

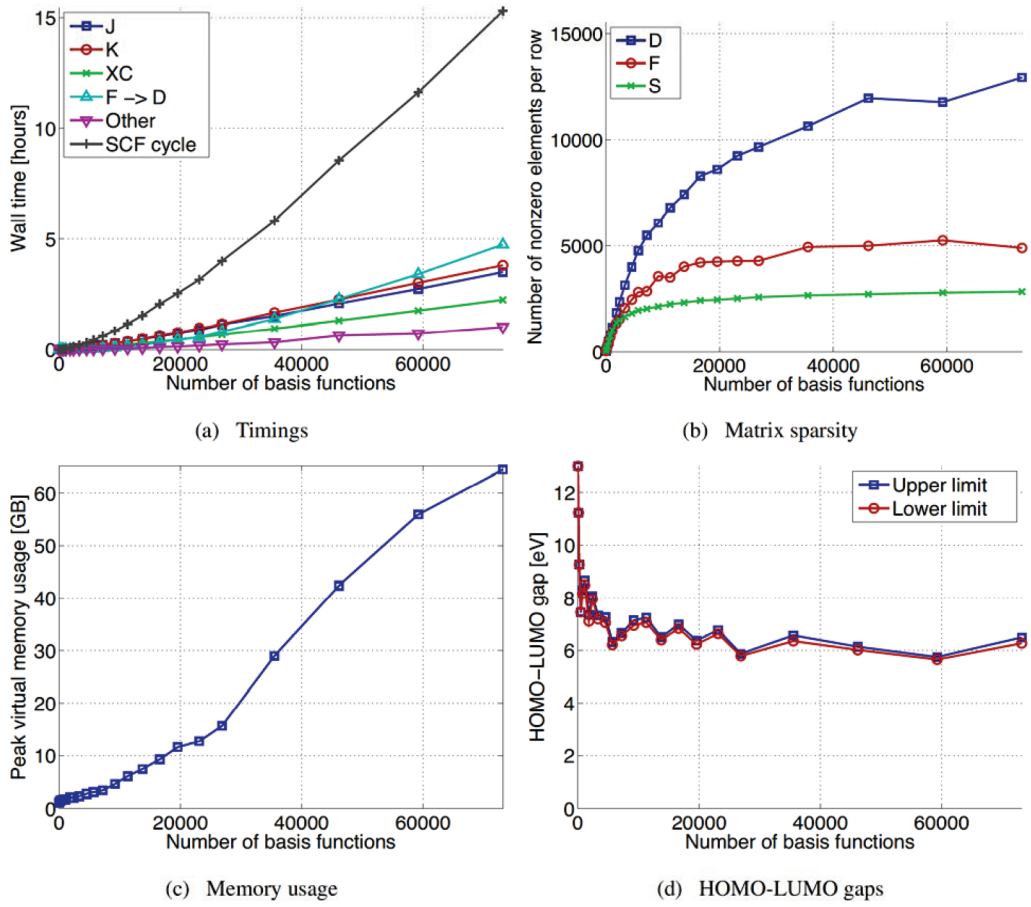


Figure 5. Timings, memory usage, matrix sparsity, and computed HOMO−LUMO gaps for BHandHLYP/6-31G** calculations on water cluster systems of varying size. The right−most points in the graphs are for a water cluster containing 3050 water molecules, corresponding to 73200 basis functions.

values to get balanced accuracy, we find that the Coulomb and HF exchange parts require a similar amount of time.

The benchmark systems used in the present paper are approximately four times larger than the systems of our previous linear scaling HF study,[4] while employing the same basis sets. This increase has been made possible not only because of greater available computer memory but also to a large extent thanks to improvements in the code. The memory usage has been significantly reduced by changing to more economic data structures and avoiding unnecessary matrix copies. Also, we have had to improve the scaling of other previously negligible parts of the code, including the computation of the overlap matrix and preparatory steps in HF exchange and Coulomb matrix construction. The performance of the used density matrix purification method[18] has been considerably improved by use of a novel scheme for the removal of small matrix elements.[39] Other technical issues arising for larger systems include changing data types of several quantities to avoid integer overflow.

One remaining issue is the scaling of the inverse Cholesky algorithm. Whereas the inverse Cholesky operation scales linearly for the Glu−Ala calculations, the scaling for the water cluster systems appears less favorable. We did not pay much attention to this issue here since the inverse Cholesky operation even for the largest water cluster calculations requires less than 5% of the total calculation time, but for larger systems, this issue is likely to become important. An alternative to the inverse Cholesky algorithm is recursive inverse factorization,[45] a method based on repeated sparse matrix multiplication.

Another important aspect is parallelization. The benchmark calculations presented in this work were performed on a single eight-core computer, using threading to exploit the eight cores. An overall speedup of around 6.5 was achieved compared to the single-core performance. The reason why the perfect speedup of eight for an eight-core machine was not reached is mainly that some parts of the code were not threaded. Clearly, this should be remedied in order to make the best use of computers with larger numbers of cores. Also, considering that many high performance computing resources are distributed memory systems, distributed memory parallelization is desirable. In future work, we aim to use a task-based approach as a way to achieve scalable parallelization of dynamic hierarchic algorithms such as the sparse matrix operations and multipole methods used in this kind of calculation.

Finally, we note a change that was needed in our density matrix purification algorithm when handling large systems. Previously, the density matrix purification scheme in the Ergo program used Theorem 3 of ref 18 to strictly ensure the correct occupation number in cases when information about the HOMO and LUMO eigenvalues is not yet available. This strict occupation number requirement has since been removed. Instead, the correct occupation number is assumed

at the end of the purification process. This change, which was necessary to handle large systems, has not caused any problems in practice. The strict occupation number requirement is in conflict with linear scaling methods where the error per electron is fixed as the system size increases. When increasing the system size, one reaches a point where the occupation number cannot be determined from the density matrix within the accuracy of one electron.

Acknowledgment. The authors thank Daniel Spångberg at Uppsala University Department of Materials Chemistry for performing the molecular dynamics simulation from which the water cluster geometries could be extracted. Support from the Swedish Research Council under Grant No. 623-2009-803 is gratefully acknowledged. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Sciences (UPPMAX) under Project p2010021.

References

(1) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.

(2) Bowler, D. R.; Miyazaki, T.; Gillan, M. J. *J. Phys.: Condens. Matter* **2002**, *14*, 2781–2798.

(3) Wu, S. Y.; Jayanthi, C. S. *Phys. Rep.* **2002**, *358*, 1–74.

(4) Rudberg, E.; Rubensson, E. H.; Sałek, P. *J. Chem. Phys.* **2008**, *128*, 184106.

(5) Hine, N.; Haynes, P.; Mostofi, A.; Skylaris, C.-K.; Payne, M. *Comput. Phys. Commun.* **2009**, *180*, 1041.

(6) Bowler, D. R.; Miyazaki, T. *J. Phys.: Condens. Matter* **2010**, *22*, 074207.

(7) White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1994**, *101*, 6593.

(8) Challacombe, M.; Schwegler, E.; Almlöf, J. *J. Chem. Phys.* **1995**, *104*, 4685–4698.

(9) Rudberg, E.; Sałek, P. *J. Chem. Phys.* **2006**, *125*, 084106.

(10) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*, 2726.

(11) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 1663–1669.

(12) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *257*, 213–223.

(13) Challacombe, M. *J. Chem. Phys.* **2000**, *113*, 10037.

(14) Li, X.-P.; Nunes, R. W.; Vanderbilt, D. *Phys. Rev. B* **1993**, *47*, 10891–10894.

(15) Goedecker, S.; Colombo, L. *Phys. Rev. Lett.* **1994**, *73*, 122–125.

(16) Palser, A. H. R.; Manolopoulos, D. E. *Phys. Rev. B* **1998**, *58*, 12704–12711.

(17) Niklasson, A. M. N. *Phys. Rev. B* **2002**, *66*, 155115.

(18) Rubensson, E. H.; Rudberg, E.; Sałek, P. *J. Chem. Phys.* **2008**, *128*, 074106.

(19) Umeda, H.; Inadomi, Y.; Watanabe, T.; Yagi, T.; Ishimoto, T.; Ikegami, T.; Tadano, H.; Sakurai, T.; Nagashima, U. *J. Comput. Chem.* **2010**, *31*, 2381.

(20) Rudberg, E.; Rubensson, E. H.; Sałek, P. *Ergo*, version 2.1. http://www.ergoscf.org (accessed Dec. 9, 2010).

(21) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular electronic-structure theory*; Wiley: Chichester, U. K., 2000.

(22) Thøgersen, L. Optimization of densities in Hartree-Fock and density-functional theory, atomic orbital based response theory, and benchmarking for radicals. Ph.D. thesis, Department of Chemistry, University of Aarhus, Aarhus, Denmark, 2005.

(23) Kudin, K. N.; Scuseria, G. E. *Math. Model. Num. Anal.* **2007**, *41*, 281–296.

(24) Zerner, M. C.; Hehenberger, M. *Chem. Phys. Lett.* **1979**, *62*, 550–554.

(25) Cancès, E.; Le Bris, C. *Int. J. Quantum Chem.* **2000**, *79*, 82–90.

(26) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393.

(27) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556.

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

(29) Panas, I.; Almlöf, J.; Feyereisen, M. W. *Int. J. Quantum Chem.* **1991**, *40*, 797–807.

(30) Panas, I.; Almlöf, J. *Int. J. Quantum Chem.* **1992**, *42*, 1073–1089.

(31) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1994**, *230*, 8–16.

(32) Schwegler, E.; Challacombe, M.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 8764–8769.

(33) Burant, J. C.; Scuseria, G. E. *J. Chem. Phys.* **1996**, *105*, 8969.

(34) Schwegler, E.; Challacombe, M.; Head-Gordon, M. *J. Chem. Phys.* **1997**, *106*, 9708.

(35) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1999**, *111*, 6223.

(36) Ochsenfeld, C. *Chem. Phys. Lett.* **2000**, *327*, 216.

(37) Lambrecht, D. S.; Ochsenfeld, C. *J. Chem. Phys.* **2005**, *123*, 184101.

(38) Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, 194106.

(39) Rubensson, E. H.; Rudberg, E. *J. Comput. Chem.* **2010**, in press.

(40) Millam, J. M.; Scuseria, G. E. *J. Chem. Phys.* **1997**, *106*, 5569–5577.

(41) Benzi, M.; Kouhia, R.; Tuma, M. *Comput. Methods Appl. Mech. Eng.* **2001**, *190*, 6533–6554.

(42) Rubensson, E. H.; Rudberg, E.; Sałek, P. *J. Comput. Chem.* **2007**, *28*, 2531–2537.

(43) Niklasson, A. M. N. *Phys. Rev. B* **2004**, *70*, 193102.

(44) Jansík, B.; Høst, S.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2007**, *126*, 124104.

(45) Rubensson, E. H.; Bock, N.; Holmström, E.; Niklasson, A. M. N. *J. Chem. Phys.* **2008**, *128*, 104105.

(46) Kohn, W.; Sham, L. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(47) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346–354.

(48) Lebedev, V. I. *Zh. Vychisl. Mat. Mat. Fiz.* **1975**, *45*, 48–54.

(49) Murray, C. W.; Handy, N. C.; Laming, G. J. *Mol. Phys.* **1993**, *78*, 997.

(50) Lindh, R.; Malmqvist, P.-Å.; Gagliardi, L. *Theor. Chem. Acc.* **2001**, *106*, 178–187.

(51) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 2547.

(52) Pissanetsky, S. *Sparse Matrix Technology*; Academic Press: New York, 1984.

(53) *Spartan ′02*; Wavefunction, Inc.: Irvine, CA, 2002.

(54) Rubensson, E. H.; Rudberg, E.; Sałek, P. *J. Math. Phys.* **2008**, *49*, 032103.

(55) Rubensson, E. H.; Zahedi, S. *J. Chem. Phys.* **2008**, *128*, 176101.

(56) Rudberg, E.; Rubensson, E. H.; Sałek, P. *J. Chem. Theory Comput.* **2009**, *5*, 80–85.

(57) Goto, K.; van de Geijn, R. A. *ACM Trans. Math. Software* **2008**, *34*, 12.

(58) GotoBLAS2. http://www.tacc.utexas.edu/tacc-projects/gotoblas2 (accessed Jan 21, 2010).

(59) Rubensson, E. H.; Rudberg, E.; Sałek, P. *Proc. PARA'06, Springer LNCS* **2007**, *4699*, 90–99.

(60) Rubensson, E. H.; Rudberg, E.; Sałek, P. *J. Comput. Chem.* **2009**, *30*, 974–977.

CT100611Z

JCTC Journal of Chemical Theory and Computation

Fast Sparse Cholesky Decomposition and Inversion using Nested Dissection Matrix Reordering

Kai Brandhorst*,† and Martin Head-Gordon*,†,‡

Department of Chemistry, University of California, Berkeley, California 94720, United States, and Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States

Received October 29, 2010

Abstract: Here we present an efficient, yet nonlinear scaling, algorithm for the computation of Cholesky factors of sparse symmetric positive definite matrices and their inverses. The key feature of this implementation is the separation of the task into an algebraic and a numeric part. The algebraic part of the algorithm attempts to find a reordering of the rows and columns which preserves at least some degree of sparsity and afterward determines the exact nonzero structure of both the Cholesky factor and its corresponding inverse. It is based on graph theory and does not involve any kind of numerical thresholding. This preprocessing then allows for a very efficient implementation of the numerical factorization step. Furthermore this approach even allows use of highly optimized dense linear algebra kernels which leads to yet another performance boost. We will show some illustrative timings of our sparse code and compare it to the standard library implementation and a recent sparse implementation using thresholding. We conclude with some comments on how to deal with positive semidefinite matrices.

1. Introduction

The Cholesky factorization has become increasingly significant in quantum chemistry, especially with respect to applications where large sparse matrices occur. This can be attributed to the fact that a Cholesky factor of a sparse symmetric positive definite matrix usually retains some degree of sparsity that can be exploited in low-order scaling algorithms.

Beebe and Linderberg[1] were probably the first who utilized the Cholesky factorization of the two-electron integral matrix in order to achieve savings in computational time for its generation and transformation. After this seminal work other groups adopted this approach,[2] and it has proven to be useful for factorizing overlap[3,4] and density matrices[5,6] in order to generate sparse transformation matrices. Furthermore, Cholesky factorizations have been used for coordinate transformations,[7] factorization of the amplitude matrix in scaled-opposite-spin MP2 (SOS-MP2)[8,9] and generation of auxiliary basis sets.[10]

While the computational savings gained by factorizing two-electron integral and density matrices stem from the fact that both are semidefinite and thus their Cholesky factors have less columns than rows, overlap matrices are strictly positive definite as long as there are no linear dependencies among the basis functions, and computational savings have been obtained by preserving the sparsity during the factorization. This approach is particularly useful for density matrix-based schemes[3,11−14] where transformations between co- and contravariant quantities are necessary.[3,13,15] For these transformations, the inverse of the metric, i.e., the inverse overlap matrix, is required, which in a dense implementation scales cubically with respect to the size of the matrix.

As two-center overlap matrices in an atomic orbital base tend to become very sparse in the large molecule limit, by contrast to their inverses, their Cholesky factors (or square roots)[16] may retain at least some degree of sparsity. It has already been pointed out[3] that this factorization can be done very efficiently by exploiting sparsity, and recently the group of Ochsenfeld[4] has devised an algorithm which is able to

* Corresponding author. E-mail: k.brandhorst@berkeley.edu; mhg@cchem.berkeley.edu.

† Department of Chemistry, University of California.

‡ Chemical Sciences Division, Lawrence Berkeley National Laboratory.

10.1021/ct100618s
Published on Web 01/10/2011

compute a sparse Cholesky factorization and its inverse by neglecting values that fall below a specified threshold, which they claim to be asymptotically linear scaling. Although this algorithm performs quite well for very sparse matrices, the performance of routines that exploit sparsity is of utmost importance in that respect, that the crossover in runtime between them and their usually highly optimized dense equivalents has to occur reasonably early in order to be advantageous. The aim of this report will be to present a more precise and yet more efficient algorithm that relies on purely algebraic methods based on graph theory, which allows the prediction of the exact nonzero structure of the Cholesky factor and its inverse before the actual numerical factorization is started. By avoiding numerical thresholding, we are even able to employ highly optimized linear algebra kernels from the BLAS[17] and LAPACK[18] libraries. We have implemented all of these improvements in a library and included it into a developer version of Q-CHEM.[19]

We want to stress that this report is more or less a brief introductory review of established mathematical methods[20] and efficient libraries for sparse Cholesky factorization have already been developed (e.g., CHOLMOD[21] and PARDISO).[22] However, none of these libraries is able to compute a sparse inverse of a Cholesky factor which is often required in quantum chemical calculations, e.g., for the transformation between orthogonalized and regular atomic orbital basis sets.[3,12,23]

2. Theory

A symmetric $N \times N$ matrix $\mathbf{A}$ is positive definite if all its eigenvalues are positive. In the case that some of the eigenvalues are zero, the matrix is positive semidefinite. Every positive semidefinite matrix $\mathbf{A}$ can be decomposed into the form $\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, with $\mathbf{X}$ having full rank if $\mathbf{A}$ is positive definite and reduced rank if $\mathbf{A}$ is positive semidefinite.[24] Among the infinite number of possible matrices $\mathbf{X}$, however, there exists a unique triangular matrix $\mathbf{L}$. This particular matrix is called the Cholesky factor[24] of $\mathbf{A}$, and its elements $L_{i,j}$ are algebraically given by:

$$L_{i,j} = \begin{cases} \frac{1}{L_{j,j}}\left(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k}L_{j,k}\right), & i > j,\ L_{j,j} \neq 0 \\ 0, & i > j,\ L_{j,j} = 0 \\ \sqrt{A_{i,i} - \sum_{k=1}^{i-1} L_{i,k}^2}, & i = j \\ 0, & i < j \end{cases} \quad (1)$$

We want to stress that the case $L_{j,j} = 0$ only happens if the matrix to be composed is semidefinite, and although the given formula algebraically holds true for any positive (semi)definite matrix $\mathbf{A}$, it is not recommended to use it for factorizing semidefinite matrices in the presence of rounding error.[25] This is due to the fact that rounding errors can accumulate and lead to almost arbitrary results. One thus has to use pivoting techniques in order to get meaningful results,[25] and we will come back to this point later. For now we will assume the matrix $\mathbf{A}$ to be strictly positive definite.

Most often the Cholesky factorization is used for solving linear sets of equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{y} \quad (2)$$

where $\mathbf{A}$ is a positive definite coefficient matrix, $\mathbf{y}$ denotes one or more right-hand side vectors, and $\mathbf{x}$ is the solution to be determined. While the naïve solution to this problem can be found by multiplying eq 2 from the left by $\mathbf{A}^{-1}$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} \quad (3)$$

and thus requires the explicit formation of the inverse of $\mathbf{A}$, this way is not recommended for practical applications for several reasons: The computation of the inverse is a quite expensive task from a computational point of view. The inverse is usually completely dense[26] even if $\mathbf{A}$ is sparse, and this approach is not backward stable[27] and can thus introduce rounding errors.

For a strictly positive definite matrix $\mathbf{A}$, however, the computationally more efficient way[28] for solving eq 2 consists of computing the Cholesky factorization of $\mathbf{A}$

$$\mathbf{L}\mathbf{L}^{\mathrm{T}}\mathbf{x} = \mathbf{y} \quad (4)$$

and finding the solution by forward and back substitution.[29] This approach is far superior to computing the inverse $\mathbf{A}^{-1}$ explicitly since computing the Cholesky factor is much less demanding. (The most efficient way of computing an inverse of a symmetric positive definite matrix is by computing its Cholesky factor, inverting it, and forming $\mathbf{A}^{-1} = \mathbf{L}^{\mathrm{T}-1}\mathbf{L}^{-1}$.) Furthermore, this approach is backward stable, and thus the solution obtained by forward and back substitution is usually more accurate when computed in the presence of rounding error. The most appealing feature of this approach however is the fact, that by contrast to inverses, Cholesky factors usually retain some degree of sparsity if the matrix to be decomposed is already sparse.

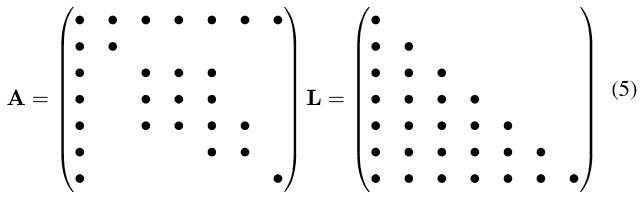
3. Sparse Cholesky Factorization

As stated above, Cholesky factors of sparse matrices tend to remain quite sparse, although they are usually not as sparse. This is due to the effect of fill-in, i.e., some elements that have been zero in the symmetric matrix become nonzero in the Cholesky factor.

3.1. Fill-in. In order to understand fill-in we need to take a closer look at eq 1. Column j of the Cholesky factor depends on the elements of all previous columns of the Cholesky factor, since these terms appear in the sum $\Sigma_{k=1}^{j-1} L_{i,k}L_{j,k}$. Note that instead of performing this summation in a single step right before the factorization of column j, we could also have performed a rank update to the right after the factorization of all columns 1, ..., $j - 1$. If the columns are factorized one at a time, then the rank-1 update is just the product of the vector with elements $(L_{i+1,j}, ..., L_{N,j})$ times its transpose. This product results in a matrix which has to be subtracted from the lower right-hand submatrix of $\mathbf{A}$.

To illustrate this effect consider the schematic representation of a sparse matrix $\mathbf{A}$ and its Cholesky factor $\mathbf{L}$. Since the algebraic nonzero structure of the Cholesky factor does not depend on the actual numerical values of $\mathbf{A}$, we simply use the symbol • to indicate nonzeros:

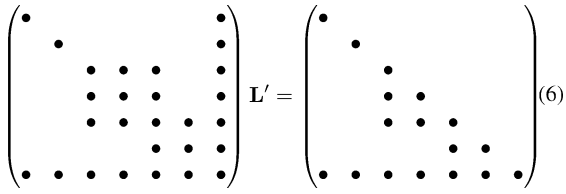
$$\mathbf{A} = \quad \mathbf{L} = \quad (5)$$



Although **A** has at least some degree of sparsity, its Cholesky factor **L** is completely dense. This is due to the rank update applied after having factorized the first column. Since the first column of **A** is dense, the product of the first column without the diagonal element times its transpose results in a completely dense matrix which has to be subtracted from the lower right-hand side submatrix of **A**. Thus the factorization continues on a completely dense matrix, resulting in a dense Cholesky factor **L**.

As a second example consider the Cholesky factorization of the matrix **A′**:

$$\mathbf{A}' = \quad \mathbf{L}' = \quad (6)$$



Now the Cholesky factor **L′** has the same nonzero structure as the matrix **A′**. As the reader can verify, all matrices used for the rank updates have nonzeros in positions that are already nonzero in **A′**, and thus no fill-in occurs at all.

From these two simple examples it becomes clear that the degree of sparsity in the Cholesky factor not only depends on the number of nonzeros in the matrix being decomposed but also strongly depends on the nonzero pattern.

3.2. Permutation Matrices. Of course the matrices **A** and **A′** are quite similar. In fact they are related by the unitary transformation

$$\mathbf{A}' = \mathbf{PAP}^{\mathbf{T}} \quad (7)$$

with

$$\mathbf{P} = \begin{pmatrix} & & & & & & 1 \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ 1 & & & & & & \end{pmatrix} \quad (8)$$

Matrices like **P** are usually called permutation matrices, since if they are applied to a matrix **A** according to eq 7, then they effectively permute rows and columns. Permutation matrices themselves can be generated by exchanging rows or columns of the identity matrix **I**.

We have already seen that the Cholesky factor **L′** of **A′** does not suffer from any fill-in, while on the other hand **L** is completely dense. Now suppose that we want to solve the linear set of eq 2 with matrix **A** from eq 5 via a Cholesky factorization and a forward/back substitution. Obviously, we cannot take advantage of sparsity in the Cholesky factors right away. However, due to $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$, we can reformulate the set of equations:

$$\mathbf{P}^{\mathrm{T}}\mathbf{PAP}^{\mathrm{T}}\mathbf{Px} = \mathbf{y} \quad (9)$$

Since eq 9 is exactly equal to eq 2, one can solve the equivalent linear set of equations

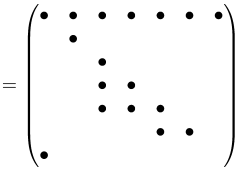
$$\mathbf{PAP}^{\mathrm{T}}\mathbf{Px} = \mathbf{Py} \quad (10)$$

$$\mathbf{A}'\mathbf{x}' = \mathbf{y}' \quad (11)$$

by applying the permutation to the right-hand side and computing the Cholesky factorization $\mathbf{A}' = \mathbf{L}'\mathbf{L}'^{\mathrm{T}}$ instead. This now allows to take full advantage of the sparsity of **A′**, since **L′** does not suffer from any fill-in during its generation. Of course we now obtain a different solution **x′**, however, the solution **x** can easily be derived by unapplying the permutation $\mathbf{x} = \mathbf{P}^{\mathrm{T}}\mathbf{x}'$.

Our focus however, is not the efficient solution of linear sets of equations but the computation of sparse Cholesky factors and their inverses, and of course **L′** is not the Cholesky factor of **A**, nevertheless

$$\mathbf{P}^{\mathrm{T}}\mathbf{L}' = \quad (12)$$

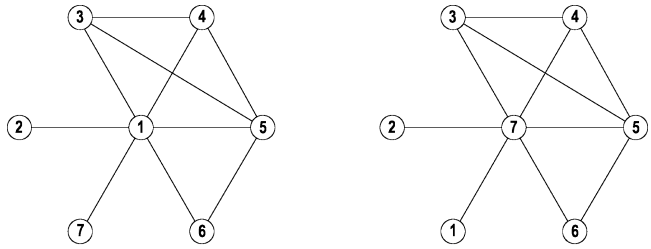


is a sparse, though nontriangular, matrix **X** which fulfills $\mathbf{A} = \mathbf{XX}^{\mathrm{T}}$. Often, this property is sufficient for **X** being useful as a transformation matrix, and it is more important that it is as sparse as possible.

As shown in this simple example, choosing an appropriate permutation matrix **P** can have a dramatic effect on the amount of fill-in occurring during the Cholesky factorization, and we will now outline how such permutation matrices can be found. We would like to stress that we have not made any assumptions on the actual numerical entries of **A** other than that the matrix be positive definite, and we will continue to do so. In what follows we will illustrate that knowing the nonzero structure of a sparse symmetric positive matrix **A** suffices to determine a permutation that results in a fairly low amount of fill-in during the factorization.

3.3. Symbolic Cholesky Factorization. The task of finding a fill-reducing permutation matrix **P** and the prediction of the exact nonzero pattern of a Cholesky factor is commonly termed *symbolic Cholesky factorization*,[20,30] since all these steps can be carried out by using purely algebraic methods from graph theory. Knowing the precise structure of Cholesky factors enables the design of efficient codes for their computation, since this allows allocation of only the actually required amount of memory which can even be done in a single step, and expensive numerical thresholding is not required at all. Furthermore, by using the so-called *supernode*[21,31,32] technique, we will show how highly optimized dense level 3 BLAS[17] and LAPACK[18] kernels can be employed even for the factorization of sparse matrices.

3.3.1. Graph Theory. Undirected graphs are useful tools in the study of symmetric matrices. Any given sparse



(a) Graph for matrix **A** (b) Graph for matrix **A′**

Figure 1. Connectivity graphs for the symmetric matrices before and after applying a permutation.

symmetric matrix **A** can be structurally represented by its associated adjacency graph $G(\mathbf{A}) = [X(\mathbf{A}), E(\mathbf{A})]$, where $X(\mathbf{A}) = \{1, 2, ..., N\}$ is the set of nodes corresponding to rows and columns of the matrix, and $E(\mathbf{A}) = \{\{X_i, X_j\}, ...\}$ is the set of edges corresponding to nonzero entries.[33] In Figure 1 the corresponding graphs for the above-mentioned matrices **A** and **A′** are depicted.

In terms of graph theory, a symmetric permutation does not affect the structure of the graph but only changes the numbering of the nodes in the graph, and all adjacency graphs for any symmetric permutation of a sparse symmetric matrix are isomorphic. However, the nonzero pattern of the Cholesky factors **L** and **L′** are very different, and in fact it is the numbering of the nodes in the adjacency graph that determines the amount of fill-in occurring during the factorization.

Recall from eq 1 that the entries of a particular column in the Cholesky factor depend on the columns to its left already being factorized. That means we have to start the factorization beginning with column 1. The relationship between the graph theoretical representation and the factorization is now that the calculation of the elements of a given column corresponds to removing the corresponding node from the adjacency graph. However, after having factorized a column, a rank update to all subsequent columns has to be performed, which as we have already seen might result in additional fill-in. Since every off-diagonal nonzero is represented by an edge in the adjacency graph, upon removal of a node, some additional edges may have to be added to the graph. In graph theoretical terms, all nodes that have been connected to the removed node have to form a *clique* afterward, i.e., all these nodes have to be connected to each other, which thus might require additional edges to be introduced.

As an example consider the graph in Figure 1a. The first column of **L** is completely dense since node 1 is connected to every other node in that graph. Upon removing node 1 from this graph, one has to add all possible missing edges between all remaining nodes to form the clique, i.e. after eliminating the first node the resulting graph is fully connected and thus the resulting Cholesky factor is completely dense.

If on the other hand one applies the same elimination procedure to the graph of matrix **A′** (Figure 1b), then one can easily verify that no additional edges have to be introduced at all, thus the resulting Cholesky factor does not suffer from any fill-in and has the same nonzero structure as **A′**.

3.3.2. Finding Fill-Reducing Reorderings. Since all adjacency graphs for any symmetric permutation applied to a sparse symmetric matrix are isomorphic and the amount of fill-in only depends on the numbering of the nodes, finding a permutation that results in a small amount of fill-in is equivalent to determining an appropriate elimination sequence.

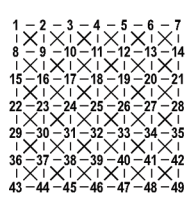
For the special case that the adjacency graph of **A** is a tree (i.e., there are no cycles in the graph), there always exists a reordering that does not introduce any additional fill-in. More generally, any chordal graph (i.e., a graph where every cycle of length at least four has an edge that connects two nonconsecutive nodes on the cycle) has a perfect elimination ordering.[33] The adjacency graph, e.g., of matrix **A** is chordal. While linear time implementations exist[34] that can test whether or not a given graph is chordal, the adjacency graphs of most matrices, however, do not fall into this category. Nevertheless a good fill-reducing reordering may still exist.

Unfortunately the task of finding an optimal, i.e., least fill-in, reordering is known to be NP-hard[35] and is thus not feasible. However, several methods have been established that provide low fill-in reorderings in polynomial time by using heuristics.

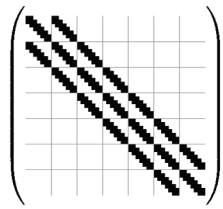
Among the most popular ones of these are the Reverse Cuthill−McKee (RCM),[36] the Lexicographic Breadth First Search (LexBFS),[34,37] the Minimum-Degree (MD),[28,38] (or approximate derivations (AMD) thereof),[39] and Nested Dissection (ND)[30,40−43] algorithms. The first two mentioned strategies have been implemented as $\mathcal{O}(|\mathbf{A}|)$ algorithms, i.e., their runtime is bound from above to be proportional to the number of nonzeros $|\mathbf{A}|$ in the matrix, and we would like to stress that RCM has already been applied in a chemically motivated context,[44] where it has been applied to the connectivity matrix of large molecules in order to reduce its bandwidth.

Strictly speaking, MD is an $\mathcal{O}(N^3)$ algorithm,[45] however it needs this time only for dense matrices and much less if the matrix is sparse. For ND, except for special cases,[30] no strict runtime bound has been established yet. Nevertheless, we used the METIS library[43] for our implementation, and we found that it can produce high-quality reorderings in reasonable amounts of time. Although its runtime is far higher than any other mentioned reordering strategy, this increased demand is itself far outweighed by the gains achieved during the subsequent numerical factorization. We are not going to describe the ND algorithm in great detail and rather refer the interested reader to consult the orginal research papers.[30,40−43]

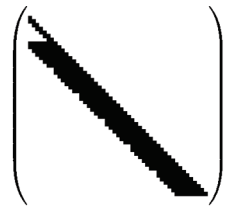
Just briefly, ND is a divide and conquer algorithm that tries to find separator nodes in a graph, i.e., nodes which upon removal would let the graph fall apart into two or more disconnected subgraphs of similar size. Those nodes are then assigned the highest node labels, i.e., they will be eliminated last. The algorithm is then applied recursively to the subgraphs until all nodes have been labeled. In Figure 2 the nonzero structures of matrices for the prominent example[46] of a regular 7×7 grid are depicted. We have used a color code in order to



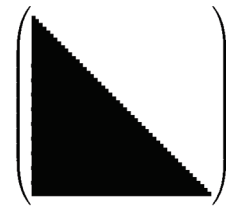
(a) Adjacency graph of a regular 7×7 grid.



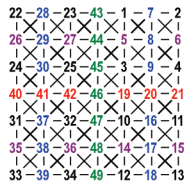
(b) Non-zero pattern of the matrix **A** corresponding to the adjacency graph in Figure 2(a)



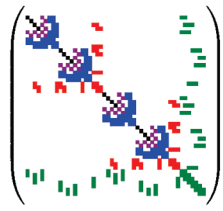
(c) Non-zero pattern of the Cholesky factor **L** of **A**



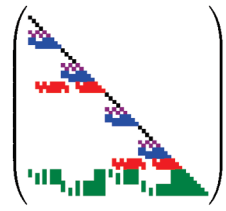
(d) Non-zero pattern of the inverse $\mathbf{L}^{-1}$ of the Cholesky factor



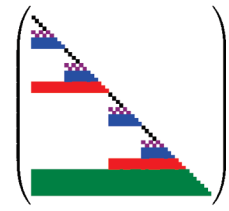
(e) Adjacency graph of a regular 7×7 grid after having applied a Nested Dissection reordering.



(f) Non-zero pattern of the matrix **A** corresponding to the adjacency graph in Figure 2(e)



(g) Non-zero pattern of the Cholesky factor $\mathbf{L}'$ of $\mathbf{A}'$



(h) Non-zero pattern of the inverse $\mathbf{L}'^{-1}$ of the Cholesky factor

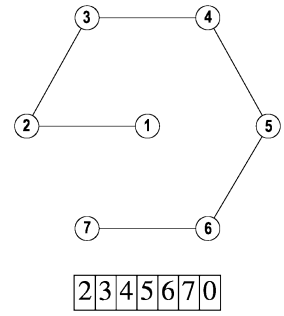
Figure 2. Illustration of a ND reordering applied to a 7×7 regular grid. (a−d) refer to matrices corresponding to the original numbering scheme, whereas (e−h) depict the matrices after applying a ND reordering. We have used a color code to highlight the individual sets of separator nodes at each level of dissection. Note the recurring patterns in (f−h).

highlight the effect of the individual steps during the ND. Matching colors indicate separator nodes of the same level, e.g., the first set of node separators is colored green. Upon their removal, the graph is separated into two disjoint subgraphs, which by removing the red nodes are further subdivided and so on and so forth. The resulting matrix $\mathbf{A}'$ (Figure 2f) shows a very characteristic pattern of recurring nonzero patterns, and these patterns will turn out to be most important for the computation of the inverse of the Cholesky factor.

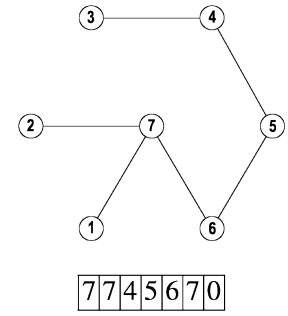
In terms of their sparsity both matrices **L** and $\mathbf{L}'$ are quite similar. This is due to the fact that the initial ordering is already quite good and is similar to a RCM ordering. In fact, all mentioned heuristics are able to find good fill-in reducing orderings, and in terms of the number of nonzeros in the Cholesky factor, they are more or less equivalent. ND, however, has one distinctive feature which makes it by far the most useful strategy if one is also interested in computing the inverse of the Cholesky factor. However, before we can explain the reasons for this, it is necessary to introduce the concept of elimination trees.[33]

3.3.3. Elimination Trees. The elimination tree[33] is defined as the adjacency graph of the Cholesky factor **L** from which all nonzeros below the diagonal except for the first one of each column have been removed, as indicated by °. For instance, referring back to the examples of eqs 5 and 6 we have

$$\mathbf{L} = \begin{pmatrix} \bullet & & & & & & \\ \bullet & \bullet & & & & & \\ \circ & \bullet & \bullet & & & & \\ \circ & \circ & \bullet & \bullet & & & \\ \circ & \circ & \circ & \bullet & \bullet & & \\ \circ & \circ & \circ & \circ & \bullet & \bullet & \\ \circ & \circ & \circ & \circ & \circ & \bullet & \bullet \end{pmatrix} \quad \mathbf{L}' = \begin{pmatrix} \bullet & & & & & & \\ & \bullet & & & & & \\ & & \bullet & & & & \\ & & \circ & \bullet & \bullet & & \\ & & & \circ & \bullet & \bullet & \\ & & & & & \bullet & \bullet \\ \bullet & \bullet & \circ & \circ & \circ & \bullet & \bullet \end{pmatrix} \quad (13)$$



(a) Elimination tree for matrix **A**



(b) Elimination tree for matrix $\mathbf{A}'$

Figure 3. Elimination trees and their corresponding vectorial representations for the symmetric matrices before and after applying a permutation.

Elimination trees always have a root, which is the highest numbered node. A node which is directly connected to the root is a child node of the root, while the root is its parent. Every node has only one parent but can have several children. Nodes which do not have any children are called leaf nodes. Every elimination tree has at least one leaf node but can have several.

In Figure 3 the elimination trees for the matrices **A** and $\mathbf{A}'$ are depicted, and in both cases, the node with number 7 is the root. The elimination tree of **A** (Figure 3a) has only one leaf node (1), while that of $\mathbf{A}'$ (Figure 3b) has three leaf nodes (1, 2, 3). Elimination trees can easily be stored in the form of a parent vector, i.e., for each node its corresponding parent node is stored in an array of size *N*, as also illustrated in Figure 3.

Given the adjacency structure of a sparse matrix and an elimination sequence (i.e., a reordering of the rows and columns), the elimination tree can be determined from the actual graph as outlined in greater detail by Liu.[47] We will not describe the actual algorithm here but only mention its runtime bounds. The original algorithm presented by Liu using *path compression* has a runtime complexity of $\mathcal{O}(|\mathbf{A}| \log_2 N)$,[48] where $|\mathbf{A}|$ stands for the number of nonzeros in the matrix subject to factorization. He mentions that another version of this algorithm exists that uses *path compression and balancing*[49−51] having a lower runtime bound of $\mathcal{O}(|\mathbf{A}| \alpha(N, |\mathbf{A}|))$ with $\alpha(N, |\mathbf{A}|)$ being the functional inverse[33,49−51] of the Ackermann function.[52] This implementation relies on a more sophisticated implementation of the set/union problem,[48,49,53] however it has been found[51] that the first version is much more efficient to implement, so the latter algorithm would have a runtime advantage only for very large graphs. Strictly speaking, none of these algorithms is truly linear, but the second one is usually considered as being *almost linear*, since α can essentially be regarded as a constant for all practically relevant integers N and $|\mathbf{A}|$.

Having determined the elimination tree for a particular reordering of a sparse symmetric matrix, the next step consists of finding what is called a postordering. This is a particular form of an equivalent reordering of the nodes in the elimination tree that changes neither the number of arithmetic operations for computing the Cholesky factorization nor the structure of the resulting adjacency graph. Therefore, in terms of both storage and computational costs, any postordering is as good as the original ordering. However, we may use them to take advantage of other aspects of elimination.

In a postordering, the nodes within every subtree of the elimination tree are numbered consecutively. The root of a subtree will always be labeled last among nodes in the subtree, and it turns out that given an elimination tree, a postorder numbering can be computed in linear time by a depth-first search.[54] As a byproduct from this step we also obtain a vector *depth* of length N that contains the distance of each node in the elimination tree from the root. This vector will turn out to be useful later. For the elimination trees depicted in Figure 3 however the nodes are already numbered in accordance with a postordering, with node 7 being the root in both cases.

The elimination tree contains useful information about data dependencies. The straightforward Cholesky factorization algorithm starts with eliminating the lowest-numbered node. If the nodes in the elimination tree are labeled according to a postordering, then the lowest-numbered node is always a leaf. Note however, that for elimination trees which have more than one leaf, we could have assigned any of them the lowest number. The factorization may thus start with removing any leaf node from the elimination tree, which is equivalent to computing the corresponding column of the Cholesky factor. Since leaf nodes do not have any children, all leaf nodes can even be removed at once, and algorithms exploiting this in a parallel manner have been developed[42,55,56] as well. Repeating this process of factorizing those columns that correspond to leaf nodes can continue as long as the root is present and will stop once this node has been eliminated as well.

Although there are no strict runtime bounds known for ND, its appealing property is that it produces broad elimination trees, i.e., trees which have many leaves. Thus these reorderings are especially useful for parallel factorization. But even if one does not intend to use any form of parallelism for the factorization step, the number of leaves in the elimination tree, or more precisely the distance from the root to the leaves, determines the nonzero structure of the inverse of the Cholesky factor as we will show later.

We also note that techniques exist that aim at finding elimination trees of minimum height,[56−58] and it is known that for every graph there exists a nested dissection ordering with minimal separators which produces an elimination tree of minimum height,[59] but unfortunately this problem is NP-hard[60,61] as well. However, we have made no attempt to find an elimination tree of minimum height, since the METIS reordering already produces well-balanced elimination trees of low height.

3.3.4. Nonzero Pattern of the Cholesky Factor. After having determined an elimination sequence that leads to a reasonable amount of fill-in, the concluding step of the symbolic Cholesky factorization is the determination of the actual nonzero pattern. In order to be memory efficient, we will use the compressed sparse column (CSC) storage scheme,[62] i.e., for every column we will store the row indices and the numerical data for the nonzero elements only, and while the computation of the numerical values will be postponed to the numerical factorization step, the determination of the row indices needs to be performed during the symbolic step. Note, that due to $\mathbf{A}'$ being symmetric, we actually only need to store the lower triangular part of this matrix.

However, before we can proceed to calculate the actual row indices we need to allocate the appropriate amount of memory required to store them, and thus we need to know how many nonzeros the Cholesky factor will have. It turns out that efficient algorithms for this task have already been developed,[51] and as before, we will only report on their runtime bounds instead of explaining the algorithm in detail. Actually the algorithm calculates the number of nonzeros for every row/column and is related to the algorithm employed for the determination of the elimination tree. It is thus not surprising that the computational complexity again depends on how the set/union problem[48,49,53] is implemented, and while the best known implementation has a complexity of $\mathcal{O}(|\mathbf{A}| \alpha(N, |\mathbf{A}|))$, still, the computationally more efficient implementation[51] has a slightly higher complexity of $\mathcal{O}(|\mathbf{A}| \log_2 N)$.

Knowing the exact number of nonzeros now allows us to allocate the actual storage space for the CSC structure of the Cholesky factor, and we can proceed with determining the actual row and column indices now. Since all $|\mathbf{L}|$ nonzero elements of the Cholesky factor need to be determined, the best runtime complexity that can be expected for this task is $\mathcal{O}(|\mathbf{L}|)$. Indeed an efficient and quite simple algorithm exists[63] that operates on the nonzero structure of the permuted matrix $\mathbf{A}$ and its corresponding elimination tree which meets this runtime bound.

3.3.5. Supernodes. Now that we have determined a proper fill-reducing reordering and the actual nonzero structure of the resulting Cholesky factor, we are almost ready to perform the actual numerical factorization. But before we do so, it is important to introduce the concept of supernodes.

Supernodes[21,31,32] are collections of columns in the Cholesky factor that share a similar nonzero pattern. Let x_p be a parent node of x_i in the elimination tree. Both nodes belong to the same supernode if the nonzero structure of column x_p is the same as that for x_i, except for the diagonal element of column x_i, and x_i is the only child of x_p. Since we assume the nodes in the elimination tree to be labeled according to a postordering, a supernode is thus a collection of columns in **L** that form a clique in the graph and share the same set of adjacent nodes outside the clique. Thus each supernode corresponds to a completely dense triangular diagonal block of the Cholesky factor and has the same nonzero pattern below the diagonal block. For the simple example given earlier, the supernodal partioning scheme is as follows:

$$\mathbf{L}' = \begin{pmatrix} \bullet & & & & \\ & \bullet & & & \\ & & \bullet & & \\ & & \bullet & \bullet & \\ & & \bullet & \bullet & \\ & & & \bullet & \\ & & & \bullet & \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{pmatrix} \tag{14}$$

It will turn out that knowing the supernodal partitioning scheme allows for a very efficient numerical factorization at the cost of only slightly increased memory requirements. This cost however is more than outweighed by the gains in speed one obtains, since this partitioning scheme allows to employ highly optimized dense BLAS and LAPACK kernels for the numerical factorization step even for sparse matrices.

3.4. Numeric Factorization. After having completed the symbolic Cholesky factorization, we will now outline how the numeric factorization can be carried out efficiently. As mentioned above, we want to take advantage of the supernodal partitioning scheme of the resulting sparse Cholesky factor by employing highly optimized dense BLAS and LAPACK kernels. For didactic reasons however, it is useful to briefly explain how a dense Cholesky factorization is implemented in LAPACK.

3.4.1. Level 3 BLAS Factorization. Most modern computers have cache architectures,[64] i.e., besides the main memory (RAM) they have up to three layers of memory that are located near the CPU which provide faster data access than to RAM. Accessing data in the level 1 cache is the fastest and has almost no latency, i.e., the level 1 cache can deliver data at every clock cycle of the CPU. However, cache is much more expensive than RAM, and thus the size of the level 1 cache is limited to usually a few kilobytes. By contrast, the RAM is the slowest memory, and as a rule of thumb, fetching data from RAM takes up to 20 clock cycles.[65] This is the reason why a naïve implementation of a matrix−matrix multiply will run at ∼5% of the theoretical peak performance, since ∼95% of the time the CPU is idle while waiting for data. More sophisticated implementations of basic linear algebra kernels like those in the BLAS library take this cache hierarchy into account and operate on smaller blocks of matrices that completely fit into the cache.[66] This effectively reduces the amount of memory fetches from RAM and leads to much more efficient codes which can run as fast as ∼95% of the theoretical peak performance of the CPU, i.e., they are about 20 times as fast.

The LAPACK library contains a lot of useful and highly optimized kernels that make use of this blocking technique as well and the routine DPOTRF, e.g., is the level 3 BLAS implementation of the Cholesky factorization for dense matrices. We will briefly review this algorithm now.

Let the lower triangular part of a symmetric positive definite matrix **A** be divided into six blocks A_1, ..., A_6 with the diagonal blocks A_1, A_4, A_6 being quadratic

A_1		
A_2	A_4	
A_3	A_5	A_6

then we start the factorization on the first diagonal block by using the LAPACK function DPOTRF:

L_1		
A_2	A_4	
A_3	A_5	A_6

As the reader can verify, the entries of the blocks below the diagonal one can be computed as the solution of a linear set of equations $\mathbf{y} = \mathbf{Tx}$, where **T** is a triangular matrix, namely the Cholesky factor of the diagonal block. We can thus simply use the level 3 BLAS function DTRSM for this task:

L_1		
$A_2 \cdot L_1^{\mathrm{T}-1}$	A_4	
$A_3 \cdot L_1^{\mathrm{T}-1}$	A_5	A_6

This concludes the factorization of the first column block, but before we can proceed, we have to update all remaining column blocks to the right of the current column block. This is essentially a symmetric rank update and we can use the BLAS kernel DSYRK for this task:

L_1		
L_2	$A_4 - L_2 \cdot L_2^{\mathrm{T}}$	
L_3	$A_5 - L_3 \cdot L_2^{\mathrm{T}}$	$A_6 - L_3 \cdot L_3^{\mathrm{T}}$

Now we can continue with factorizing the second diagonal block, by using DPOTRF again

L_1		
L_2	L_4	
L_3	A_5'	A_6'

and by using DTRSM we can factorize the remaining blocks of the current column:

L_1		
L_2	L_4	
L_3	$A'_5 \cdot L_4^{T-1}$	A'_6

Again, we apply the rank update to the remaining column with DSYRK

L_1		
L_2	L_4	
L_3	L_5	$A'_6 - L_5 \cdot L_5^T$

and likewise we would continue with this process until the last diagonal block has been factorized

L_1		
L_2	L_4	
L_3	L_5	L_6

which concludes the factorization.

After this brief explanation on how to compute the Cholesky factorization of a dense matrix, we will now turn to the discussion of a corresponding sparse implementation. As stated above, the collection of nodes that belong to the same supernode forms a clique in the adjacency graph of $\mathbf{L}$, and therefore, the diagonal block of each supernode is always completely dense, and we can employ the LAPACK kernel DPOTRF for its factorization. In order to be able to do so however, we have to set up an appropriate data structure that allows for the invocation of this function, i.e., we have to store the diagonal blocks as a dense matrix, and although the function DPOTRF only operates on the lower (or upper) triangular part of the matrix to be factorized, the data has still to be passed as a full $n \times n$ array, where n is the dimension of the matrix. That means, we have to allocate the upper triangular part of each diagonal block of every supernode as well as indicated by °. This additional memory

$$\mathbf{L}' = \begin{pmatrix} \bullet & & & & \\ & \bullet & & & \\ & & \bullet\ \circ & & \\ & & \bullet\ \bullet & & \\ & & \bullet\ \bullet & \bullet\ \circ\ \circ & \\ & & & \bullet\ \bullet\ \circ & \\ \bullet & \bullet & \bullet\ \bullet & \bullet\ \bullet\ \bullet & \end{pmatrix} \tag{15}$$

requirement thus causes a small overhead which however is more than outweighed by the increased performance compared to a sparse indexing based factorization.

So far of course this only allows for the efficient computation of the diagonal blocks of $\mathbf{L}'$. However, we know that all elements below the diagonal block share the same row indices and can thus, by omitting the zero rows, be stored compactly as a dense rectangular matrix as well. Note that we can actually store the whole supernode as a dense matrix if we use the LDA parameter of the BLAS and LAPACK routines properly. The proper data structure for the numerical factorization is thus essentially a collection of dense matrices (one for every supernode) that contain the numerical values of the lower triangular part of the matrix $\mathbf{A}'$, as illustrated in Figure 4.

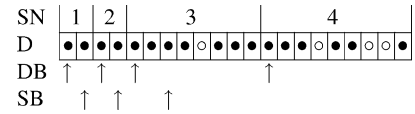


Figure 4. Illustration of the supernodal storage scheme used for the Cholesky factor $\mathbf{L}'$. SN indicates the number of the supernode, D holds the actual numerical data stored in column major order, where ° denotes the additional storage required in order to be able to invoke BLAS and LAPACK routines. DB and SB indicate the pointers to the beginning of the diagonal and subdiagonal blocks, respectively.

This storage scheme now allows us to proceed with the remaining steps of the factorization in a similar way as illustrated for the dense factorization. After having factorized the diagonal block of a supernode, we can employ the dense matrix kernel DTRSM to compute the remaining parts below the diagonal block, and since the subdiagonal part is stored as a dense matrix, we can yet use the level 3 BLAS function DSYRK for the computation of the rank update of the columns to the right of the current supernode. However, we have to store the result in a temporary array first and then scatter the data into their final destinations, since the nonzero pattern of subsequent supernodes is not necessarily the same as that of the current supernode. That means, we need one more temporary array to store the rank update. The size of this array can be determined prior to the computation since it only depends on the size of the largest subdiagonal block and is quite small for sparse matrices.

This sequence can then be carried out for every supernode and will eventually lead to the Cholesky factor of the matrix $\mathbf{A}'$. Note that the last supernode always consists of just a dense square block. The size of this block tends to become larger the denser the Cholesky factor is. Thus in case the matrix $\mathbf{A}'$ is already quite dense, there will be only one large supernode which is factorized by using the highly optimized LAPACK kernel DPOTRF. That means, we can expect a seamless transition in performance from very sparse to completely dense matrices. Although there is an additional overhead for the symbolic Cholesky factorization, that can of course be avoided if one knows that the Cholesky factor is going to be completely dense; there will be no significant loss in performance as compared to DPOTRF if the matrix is more or less dense. This behavior is probably the most important advantage of using reordering and supernode techniques for the factorization, since the performance will not be affected as much as it is often the case for sparse matrix routines if the matrix is dense.

4. Inverses of Cholesky Factors

The elements $L_{i,j}^{-1}$ of the inverted Cholesky factor $\mathbf{L}^{-1}$ are given by

$$L_{i,j}^{-1} = \begin{cases} \dfrac{1}{L_{i,i}}\left(-\sum_{k=j}^{i-1} L_{i,k} L_{k,j}^{-1}\right), & i > j \\ \dfrac{1}{L_{i,i}}, & i = j \end{cases} \tag{16}$$

As stated above, for the solution of linear sets of equations like eq 2 there is usually no need to explicitly compute $\mathbf{A}^{-1}$ or

$\mathbf{L}^{-1}$ since one can find the solution by forward/back substitution. In the presence of rounding errors, this is not only the most efficient way but also the most accurate, since the computation of $\mathbf{L}^{-1}$ is not backward stable[27] and thus can lead to unnecessary rounding errors. Furthermore, even if the right-hand side vectors $\mathbf{y}$ are sparse, the solution $\mathbf{x}$ is usually dense, due to the fact that the inverse of $\mathbf{A}$ is usually[26] dense. In light of this, it can easily be understood why all available sparse solvers[21,22,67] always operate on dense right-hand side vectors.

However, if one intends to find the solution to the forward substitution step only and the right-hand side is very sparse, then chances are good that the result remains sparse. This is due to the fact that the result of the forward substitution is given by $\mathbf{L}^{-1}\mathbf{y}$, and as long as both matrices are sufficiently sparse, the result may still be. Furthermore, if one intends to use the inverse Cholesky factor as a transformation matrix, then it becomes obvious that for reasons of efficiency it should be as sparse as possible.

Again, we assume the Cholesky factor has been computed for the permuted matrix $\mathbf{A}'$ in order to reduce the fill-in during the factorization. Then the inverse of the original (i.e., unpermuted) matrix is given by

$$\mathbf{A}^{-1} = (\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{A}\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1} \tag{17}$$

$$= \mathbf{P}^{\mathrm{T}}(\mathbf{P}\mathbf{A}\mathbf{P}^{\mathrm{T}})^{-1}\mathbf{P} \tag{18}$$

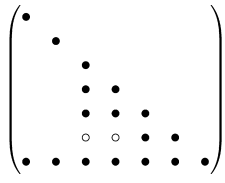
$$= \mathbf{P}^{\mathrm{T}}(\mathbf{L}'\mathbf{L}'^{\mathrm{T}})^{-1}\mathbf{P} \tag{19}$$

$$= \mathbf{P}^{\mathrm{T}}\mathbf{L}'^{-1^{\mathrm{T}}}\mathbf{L}'^{-1}\mathbf{P} \tag{20}$$

and thus a potentially sparse, though nontriangular, equivalent of the inverse of the orginal Cholesky factor is given as $\mathbf{L}'^{-1}\mathbf{P}$.

As for the Cholesky factorization, the computation of the inverse can be split up into a symbolic and a numeric part.

4.1. Symbolic Cholesky Inversion. Given the nonzero structure of the Cholesky factor, the nonzero structure of its inverse is quite easy to predict. It is defined as the *transitive closure* of the elimination tree,[26,40,61] i.e., the nonzero entries of every column are given by following the path from the corresponding node up to the root. Every node visited on the path corresponds to a nonzero element in the column. Thus for the above given example, the structure of the inverse Cholesky factor can be determined from the elimination tree depicted in Figure 3b:

$\mathbf{L}'^{-1} =$  (21)

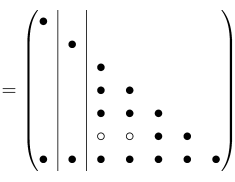
Here we use ° to indicate fill-in positions during the inversion. Fill-in usually occurs and can be avoided only in special cases (e.g., if the adjacency graph of the Cholesky factor is a tree), which means that the inverse of a Cholesky factor is usually denser than the Cholesky factor, which itself is usually denser than the matrix subject to the factorization. This is of course a huge drawback, and thus computational gains by using sparse algorithms can only be expected if the matrix is very sparse and its elimination tree is very broad. The latter point is by far the most important since the shorter the paths from the nodes to the root are the less nonzeros the inverse of the Cholesky factor will have. This observation clarifies the importance of finding fill-reducing orderings that result in an elimination tree with many leafs.

We have tested all of the above-mentioned reordering heuristics, and while all perform very well for the Cholesky factorizaton, only ND is able to produce "bushy" elimination trees, while all other reorderings tend to produce elimination trees that are more or less straight lines like the one in Figure 3a. Although ND is the most expensive reordering strategy, the quality of the reordering in terms of its supernode structure and number of nonzeros in the inverse is far superior, and the higher computational demand is by far outweighed by the savings gained during the numerical steps.

4.1.1. Nonzero Pattern of the Inverted Cholesky Factor. The symbolic part of the inversion is focused on determining the row and column indices of the nonzeros, and in order to avoid multiple memory allocation steps or a waste of memory, again we first need to know the exact number of nonzeros. Since the nonzero pattern is determined by the transitive closure of the elimination tree, the number of nonzeros in one column is given by the distance of the corresponding node from the root of the elimination tree, and as mentioned above, this information has already been computed as a byproduct of the postordering. The number of nonzeros in the whole inverse can thus easily be computed by summing over all elements in the vector *depth*, which has a complexity of $\mathcal{O}(N)$.

After having allocated the sparse data structure, we can then proceed with the determination of the actual nonzero indices. Since we have to determine all of them, the best computational complexity for this task is expected to be $\mathcal{O}(|\mathbf{L}^{-1}|)$. Since we have stored the elimination tree as a vector in which the parent of every node is stored, it is easy to verify that by following the path from a given node to the root of the elimination tree; all nonzero elements can be computed within the given runtime bound.

4.1.2. Supernodes during Inversion. We have stressed that taking into account the supernodal structure can lead to much more efficient numerical factorizations, and it turns out that we can use the same technique for the inversion as well. Due to the additional amount of fill-in, however, the supernodal structure for computing the inverse is usually different from the one used for the factorization, as illustrated for the given example:

$\mathbf{L}'^{-1} =$  (22)

The data structure that allows utilizing BLAS and LAPACK kernels is essentially the same as that used for the factorization, i.e., we store each supernode as a dense rectangular matrix.

4.2. Numeric Cholesky Inversion. The biggest difference between the computation of the Cholesky factor and the

calculation of its inverse is probably the order in which the columns have to be computed. As the reader can verify from eq 16, by contrast to the factorization, the elements of the inverse depend on elements of the inverse to their right, i.e., we have to start the inversion with the utmost right column and then proceed to the left.

As for the Cholesky factorization, the LAPACK library already contains a function DTRTRI that is able to compute the inverse of a dense triangular matrix by invoking level 3 BLAS calls, and since we are going to adopt this algorithm to the case when the triangular matrix is sparse, for didactic reasons we will illustrate the basic steps for a dense matrix.

Let the Cholesky factor **L** be divided into six blocks L_1, ..., L_6 with the diagonal blocks L_1, L_4, L_6 being quadratic

L_1		
L_2	L_4	
L_3	L_5	L_6

then we can start the factorization on the last diagonal block by using the LAPACK kernel DTRTRI:

L_1		
L_2	L_4	
L_3	L_5	L_6^{-1}

As the next step we have to update the subdiagonal block of the second column block, and as the reader can verify, this is equivalent to a matrix multiply between a triangular and a rectangular matrix. We can thus use the BLAS function DTRMM for this task:

L_1		
L_2	L_4	
L_3	$L_6^{-1} \cdot L_5$	L_6^{-1}

After this multiplication the final block L_5^{-1} can be computed as the solution to a linear set of equations with a triangular coefficient matrix by invoking the BLAS kernel DTRSM

L_1		
L_2	L_4	
L_3	$L_5' \cdot L_4^{-1}$	L_6^{-1}

before we invert the diagonal block of the second row by using DTRTRI again:

L_1		
L_2	L_4^{-1}	
L_3	L_5^{-1}	L_6^{-1}

At this point the last two columns already contain the right numerical values, and we can proceed with the first column. In principle this can be carried out by multiplying the already inverted lower right-hand submatrix with the subdiagonal block of the first column. In order to be able to do this in place, however, we have to break this up into smaller steps, and we will start by modifying block L_3 by calling DTRMM:

L_1		
L_2	L_4^{-1}	
$L_6^{-1} \cdot L_3$	L_5^{-1}	L_6^{-1}

The next step requires a matrix multiply between two rectangular matrices, which is best been done by the prominent BLAS kernel DGEMM:

L_1		
L_2	L_4^{-1}	
$L_3' + L_5^{-1} \cdot L_2$	L_5^{-1}	L_6^{-1}

Note that for this step the unmodified block L_2 is required. This is the reason why the update of L_2 by invoking DTRMM can only take place after having updated all blocks below it:

L_1		
$L_4^{-1} \cdot L_2$	L_4^{-1}	
L_3''	L_5^{-1}	L_6^{-1}

Now the multiplication by the inverse of the diagonal block of the current row can be done by using DTRSM again

L_1		
$L_2' \cdot L_1^{-1}$	L_4^{-1}	
$L_3'' \cdot L_1^{-1}$	L_5^{-1}	L_6^{-1}

and the inversion of the diagonal block L_1 with calling DTRTRI concludes the inversion:

L_1^{-1}		
L_2^{-1}	L_4^{-1}	
L_3^{-1}	L_5^{-1}	L_6^{-1}

Of course the actual sparse implementation is a little more complicated, but it is essentially identical with a slight overhead for bookkeeping. The appealing advantage in this implementation is that all operations can be done by directly calling BLAS and LAPACK functions and that there is not a single step where temporary results have to be scattered into their destinations. This and the fact that the individual supernodes are larger compared to those in the factorization step account for this algorithms' efficiency.

Table 1. CPU Timings for Computation of the Cholesky Factor **L** and Its Inverse $\mathbf{L}^{-1}$ of Two-Center Overlap Matrices **S** in the Basis Set 6-31G(d) for a Series of Linear Alkanes, Single Graphite Layers, and Spherical Diamond Blocks of Various Sizes[a]

				thresholding			supernodal			LAPACK
	n	dim(**S**)	% (**S**)	% (**L**)	% ($\mathbf{L}^{-1}$)	time (s)	% (**L′**)	% ($\mathbf{L'}^{-1}$)	time (s)	time (s)
alkanes (C_nH_{2n+2})	100	1904	8.1	11.8	30.6	2.98	9.1	16.2	0.20	0.78
	200	3804	4.1	6.8	18.5	9.41	5.1	10.5	0.48	5.64
	300	5704	2.7	4.7	13.1	16.06	3.5	8.0	0.82	18.37
	400	7604	2.1	3.6	10.1	22.76	2.7	6.5	1.19	42.31
	500	9504	1.7	2.9	8.2	29.26	2.2	5.6	1.59	82.16
	600	11404	1.4	2.4	6.9	35.12	1.8	4.8	1.96	140.58
	700	13304	1.2	2.1	6.0	42.06	1.6	4.3	2.39	220.58
	800	15204	1.0	1.8	5.2	48.18	1.4	3.9	2.82	328.05
	900	17104	0.9	1.6	4.7	56.75	1.2	3.6	3.28	465.61
	1000	19004	0.8	1.5	4.2	64.88	1.1	3.3	3.73	634.87
graphite (C_n)	100	1500	22.1	30.2	30.5	3.36	22.3	26.9	0.44	0.39
	200	3000	12.9	28.5	30.5	29.86	17.5	24.2	1.77	2.81
	300	4500	9.2	26.3	30.5	92.94	14.3	21.7	4.18	9.14
	400	6000	7.1	24.4	30.5	195.77	12.3	20.2	7.09	21.15
	500	7500	5.9	22.7	30.5	351.83	10.8	18.7	10.15	40.61
	600	9000	5.0	21.2	30.4	566.38	9.8	17.8	15.31	69.36
	700	10500	4.3	20.0	30.4	830.96	9.0	16.9	19.43	109.34
	800	12000	3.8	18.9	30.4	1156.21	8.6	16.4	27.44	162.50
	900	13500	3.4	17.9	30.4	1532.60	7.8	15.8	32.41	229.13
	1000	15000	3.1	17.1	30.4	2024.51	7.5	15.1	39.34	313.79
diamond (C_n)	100	1500	62.4	49.9	50.0	8.81	46.8	49.5	1.57	0.39
	200	3000	47.0	49.9	50.0	77.36	43.9	48.2	10.18	2.81
	300	4500	38.0	49.9	50.0	253.42	42.1	47.3	26.35	9.12
	400	6000	31.7	49.9	50.0	605.00	39.9	46.4	48.76	21.08
	500	7500	27.3	49.9	50.0	1165.08	37.9	45.2	94.72	40.65
	600	9000	24.1	49.9	50.0	2022.88	36.0	42.8	139.21	69.25
	700	10500	21.6	49.8	50.0	3231.13	35.6	43.6	226.14	109.09
	800	12000	19.5	49.7	50.0	4643.79	33.5	41.3	255.94	162.29
	900	13500	17.8	49.6	50.0	6570.94	33.1	42.2	365.23	229.99
	1000	15000	16.5	49.5	50.0	9161.83	32.0	40.4	451.80	313.92

[a] All timings are in seconds and have been carried out on a single core/single thread (`OMP_NUM_THREADS=1`) on a 2.31 GHz AMD Opteron 2376 architecture. They include memory allocation times for all sparse matrices and temporary data structures generated. The time required to calculate the overlap matrices, however, is not included. Thresholding indicates the implementation of the Ochsenfeld group, while our implementation is the supernodal. For comparison, the cumulative execution times for the LAPACK calls DPOTRF and DTRTRI as implemented in the Intel MKL are provided. Dim(**S**) stands for the dimension of the overlap matrix and % () indicates the density in percent, respectively.

5. Illustrative Timings

After having explained how the supernodal Cholesky factorization and inversion can be done, it is now time to justify our claims about its superior efficiency.

Because they tend to become very sparse in the large molecule limit and are always strictly positive definite, as long as the basis set does not contain linear dependencies among the basis functions, we choose two-center overlap matrices as our test targets. We consider linear alkanes (C_nH_{2n+2}), graphite (C_n), and diamond (C_n), with $n = 100, ..., 1000$ as representatives of one-, two- and three-dimensional geometries, respectively. For the construction of the geometries standard values were assumed for the geometrical parameters (alkanes: $r(C-C) = 1.54$ Å, $r(C-H) = 1.10$ Å, graphite: $r(C-C) = 1.42$ Å, and diamond: $r(C-C) = 1.54$ Å).

All calculations were carried out using a single core/single thread (`OMP_NUM_THREADS=1`) on a 2.31 GHz AMD Opteron 2376 architecture running linux. All timings include memory allocation times for all sparse matrices and temporary data structures generated by the routines. The computation of the overlap matrices itself, however, is not included. Both codes have been compiled with full optimization using the Intel compilers.

After having computed the overlap matrices, we disregarded all entries below a threshold of 10^{-15} and stored the remaining entries in the CSC format, which we then supplied to both our own implementation and the one of the Ochsenfeld group. (Here we used the same threshold of 10^{-15} throughout.) Since this algorithm uses thresholding criteria for retaining sparsity, we will refer to this as the "thresholding" implementation, while our own code is termed "supernodal".

The collective timings for the sparse Cholesky factorization and the subsequent inversion for the overlap matrices in the 6-31G(d) and 6-311G(2df) basis set are listed in Tables 1 and 2 and depicted in Figure 5, respectively. For comparison, we have also performed timings for the LAPACK routines DPOTRF and DTRTRI from the Intel Math Kernel (MKL) library (version 10.2).

For the linear alkanes both implementations are faster in the large molecule limit as compared to the dense LAPACK codes. This is not surprising since for both basis sets with incresing system size the density of the overlap matrix rapidly

Table 2. CPU Timings for Computation of the Cholesky Factor **L** and Its Inverse $\mathbf{L}^{-1}$ of Two-Center Overlap Matrices **S** in the Basis Set 6-311G(2df) for a Series of Linear Alkanes, Single Graphite Layers, and Spherical Diamond Blocks of Various Sizes[a]

				thresholding			supernodal			LAPACK
	n	dim(**S**)	% (**S**)	% (**L**)	% ($\mathbf{L}^{-1}$)	time (s)	% (**L′**)	% ($\mathbf{L'}^{-1}$)	time (s)	time (s)
alkanes (C_nH_{2n+2})	100	4106	6.8	10.1	31.2	27.45	8.0	15.3	1.05	7.40
	200	8206	3.5	6.2	24.4	118.46	4.4	9.9	2.76	55.14
	300	12306	2.3	4.6	20.4	260.62	3.1	7.6	4.73	178.46
	400	16406	1.7	3.6	17.6	443.93	2.4	6.2	7.05	419.51
	500	20506	1.4	3.0	15.4	644.27	1.9	5.3	9.38	811.10
	600	24606	1.2	2.6	13.7	848.45	1.6	4.7	12.12	1381.50
	700	28706	1.0	2.2	12.3	1080.14	1.4	4.1	14.63	2168.69
	800	32806	0.9	2.0	11.0	1303.58	1.2	3.8	17.65	3253.99
	900	36906	0.8	1.8	10.0	1599.31	1.1	3.4	20.43	4605.42
	1000	41006	0.7	1.6	9.2	1762.28	1.0	3.2	23.19	6304.39
graphite (C_n)	100	3500	18.0	28.2	28.5	44.19	19.8	24.5	3.12	4.41
	200	7000	10.4	27.0	28.5	331.64	14.6	21.5	13.41	33.08
	300	10500	7.4	25.1	28.5	1015.73	12.5	19.4	30.94	109.08
	400	14000	5.7	23.4	28.5	2180.21	10.8	18.0	51.01	255.74
	500	17500	4.7	21.8	28.5	3853.73	9.9	17.2	82.85	496.33
	600	21000	4.0	20.5	28.5	6181.12	8.6	16.1	114.74	852.52
	700	24500	3.4	19.3	28.5	9212.60	7.9	15.3	150.01	1352.71
	800	28000	3.0	18.3	28.4	12697.22	7.3	14.6	195.04	2017.46
	900	31500	2.7	17.4	28.4	16952.24	6.7	14.1	239.24	2864.93
	1000	35000	2.5	16.6	28.4	22187.31	6.4	13.5	296.49	3923.51
diamond (C_n)	100	3500	55.2	49.8	50.0	122.12	45.8	49.0	17.41	4.42
	200	7000	40.3	49.9	50.0	948.06	43.2	47.8	87.65	33.07
	300	10500	32.1	49.9	50.0	3143.71	39.4	44.7	268.13	108.98
	400	14000	26.6	49.9	50.0	7476.32	37.1	43.7	468.03	256.24
	500	17500	22.7	49.9	50.0	14647.72	35.5	42.7	786.33	495.17
	600	21000	20.0	49.9	50.0	24621.83	33.8	41.7	1082.80	856.80
	700	24500	17.9	49.9	50.0	38711.14	32.6	40.9	1704.99	1358.30
	800	28000	16.1	49.8	50.0	59171.17	31.2	40.0	2111.56	2025.69
	900	31500	14.7	49.8	50.0	85512.58	31.1	41.5	3548.40	2860.03
	1000	35000	13.5	49.7	50.0	115717.98	28.9	38.7	3882.37	3924.98

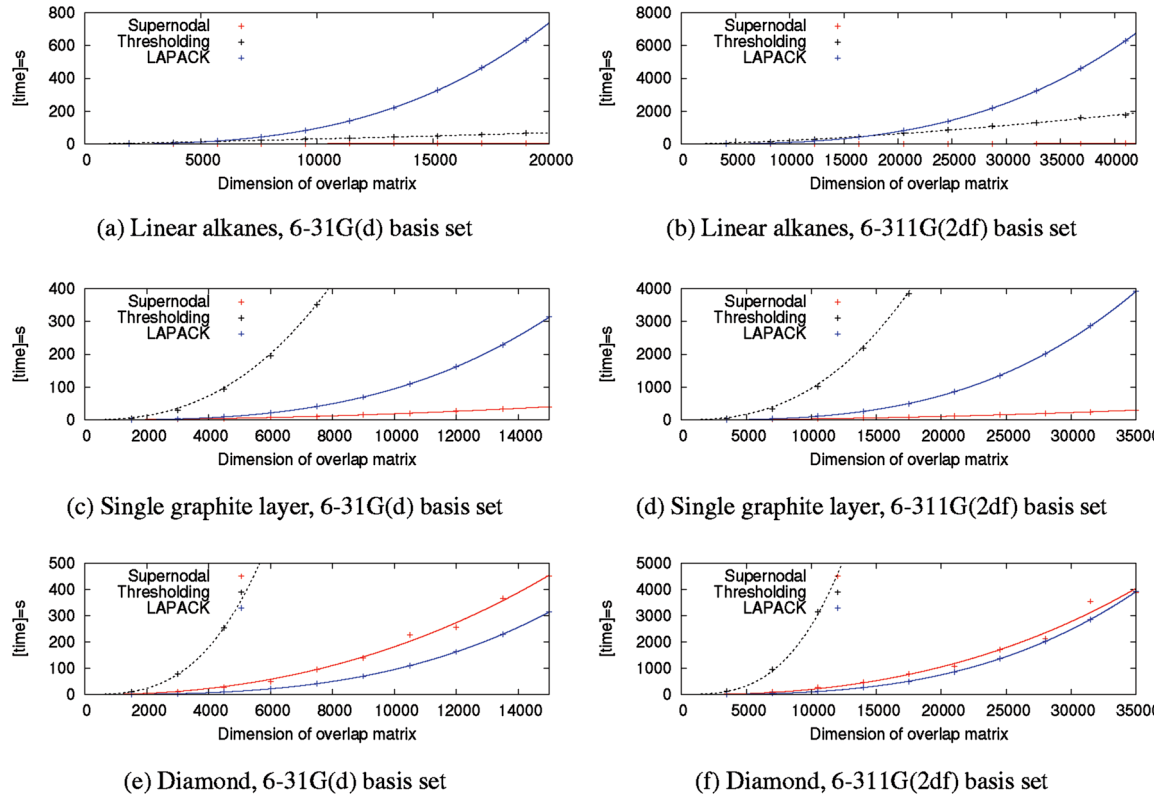
[a] All timings are in seconds and have been carried out on a single core/single thread (`OMP_NUM_THREADS=1`) on a 2.31 GHz AMD Opteron 2376 architecture. They include memory allocation times for all sparse matrices and temporary data structures generated. The time required to calculate the overlap matrices, however, is not included. Thresholding indicates the implementation of the Ochsenfeld group, while our implementation is the supernodal. For comparison, the cumulative execution times for the LAPACK calls DPOTRF and DTRTRI as implemented in the Intel MKL are provided. Dim(**S**) stands for the dimension of the overlap matrix and % () indicates the density in percent, respectively.

decays from 8.1% and 6.8% to 0.8% and 0.7%, respectively. However, while the crossover between the dense kernel and the thresholding algorithm occurs at roughly 5000 basis functions in the 6-31G(d) basis and at 17500 basis functions in the 6-311G(2df) basis, our supernodal implementation is faster than both codes throughout and runs by a factor of 14−78 faster than the thresholding implementation. This indicates that even for very sparse matrices the supernodal technique is advantageous even though the thresholding code shows a lower scaling with system size (with a larger prefactor though) according to the polynomial fit (see Table 3) for the 6-31G(d) basis. We would like to stress that strictly speaking the thresholding code as it is implemented is quadratically scaling since it requires the allocation of a full matrix in order to hold the Cholesky factor. This step of course has a tiny prefactor and could even be avoided by subsequent reallocation of memory once it is actually needed.

Since the supernodal algorithm relies on applying a fill-reducing reordering to the initial matrix, which also reduces the number of arithmetic operations, both matrices $\mathbf{L}'$ and $\mathbf{L}'^{-1}$ are less dense than their unpermuted counterparts. It is thus likely that the supernodal code will have runtime advantage for any reasonable number of basis functions and that the scaling might even further decrease for larger systems.

Of course these test systems are far from being linear in reality and should be seen as ideal test cases for the algorithms. We thus have included more realistic and less ideal test cases as well, one of them being a single graphite layer as a representative of a two-dimensional system.

For graphite the difference in performance between the two sparse implementations is even more pronounced. While the thresholding algorithm is roughly a factor of 5−10 slower than the LAPACK functions, with one exception our supernodal implementation is faster than the optimized dense kernels by up to a factor of 13. Here the superior quality of the ND reordering becomes apparent, which is able to produce Cholesky factors and their corresponding inverses, which are roughly half as dense as those obtained by the thresholding algorithm. Once again, we would like to stress that we have not applied any sort of thresholding, i.e., it is likely that the number of significant elements in the inverse could even be reduced by eliminating those values that fall below a given threshold.



(a) Linear alkanes, 6-31G(d) basis set

(b) Linear alkanes, 6-311G(2df) basis set

(c) Single graphite layer, 6-31G(d) basis set

(d) Single graphite layer, 6-311G(2df) basis set

(e) Diamond, 6-31G(d) basis set

(f) Diamond, 6-311G(2df) basis set

Figure 5. Graphical representation of the data in Tables 1 and 2.

Table 3. Coefficients *a*, *b* for the Fitting of the Data in Tables 1 and 2 According to a Fit Function of the Form $a \cdot \dim(\mathbf{S})$

		6-31G(d)		6-311G(2df)	
		a	*b*	*a*	*b*
alkanes	thresholding	6.17×10^{-4}	1.17	1.76×10^{-4}	1.52
	supernodal	1.61×10^{-5}	1.25	2.13×10^{-5}	1.31
	LAPACK	1.42×10^{-10}	2.96	1.23×10^{-10}	2.97
graphite	thresholding	6.52×10^{-8}	2.51	9.19×10^{-8}	2.50
	supernodal	5.02×10^{-7}	1.89	8.98×10^{-7}	1.87
	LAPACK	1.48×10^{-10}	2.95	1.08×10^{-10}	2.98
diamond	thresholding	3.63×10^{-9}	2.97	2.14×10^{-9}	3.02
	supernodal	1.58×10^{-7}	2.26	3.79×10^{-8}	2.43
	LAPACK	1.41×10^{-10}	2.96	1.17×10^{-10}	2.98

As a concluding challenging example we choose to test the performance of our sparse implementation on rather dense matrices, namely the overlap matrices of diamond. Of course no one would use a sparse code for a dense matrix since highly optimized BLAS and LAPACK kernels have been designed exactly for this purpose, however if the density of the matrix subject to factorization is not known in advance, then one might give a sparse implementation a try if its performance is not too bad as compared to a dense code. As can be seen from Tables 1 and 2, the thresholding algorithm is slower by a factor of 22−30 as compared to LAPACK, and the scaling is even worse. Not surprisingly, our supernodal implementation is slower than the dense equivalent as well, however it is at most 4 times slower, and by contrast to the thresholding implementation, the differences more and more vanish with increasing matrix size and even break even for the diamond cluster with 1000 atoms in the 6-311G(2df) basis. This behavior once again highlights the importance of the supernode technique, which allows for a seamless transition between really sparse systems and more or less dense ones that are treated as if they were sparse. For a completely dense matrix, our implementation would have an overhead for the symbolic Cholesky factorization. The numeric part however would be the same single LAPACK call as for the dense matrix, since the matrix would have just one supernode, and this property renders our implementation more or less generally applicable.

6. Semidefinite and Idempotent Matrices

So far we have only considered strictly positive definite matrices, and as stated above, symmetric positive semidefinite matrices do have Cholesky factors as well. However, in the presence of rounding error, one is forced to use the *full pivoting* technique[25] in order to obtain reliable results, i.e., the factorization is carried out by exchanging the current row/column with that one having the largest diagonal element among those that have not yet been factorized. This process is repeated until all remaining diagonal elements fall below a predefined threshold. The huge drawback of this approach however is that the matrix has to be reordered at every step of the factorization, and since this reordering now depends on the actual numerical values of the matrix, we cannot apply a symbolic Cholesky factorization anymore.

For semidefinite matrices which are very rank deficient (density and two-electron integral matrices, e.g.), this approach is fine. However, although overlap matrices are positive definite as long as there are no linear dependencies among the basis functions, they tend to become numerically semidefinite for larger basis sets. If this happens to be the case, then one would encounter numerical problems during the regular Cholesky factorization. Furthermore the Cholesky factor would not have a regular inverse anymore, and one

- **Prerequisites**
  - compute overlap matrix
  - disregard all entries below a defined threshold
  - store remaining significant elements in CSC format
- **Symbolic Cholesky factorization**
  - create graph structure
  - determine Nested Dissection reordering
  - determine elimination tree
  - determine postordering
  - determine number of non-zeros per row/column and supernodal structure
  - (if desired) determine number of non-zeros in inverse Cholesky factor
  - determine storage space for numerical factorization
- **Numeric Cholesky factorization**
  - create CSC structure of Cholesky factor
  - setup numerical data for supernodal factorization
  - perform numeric factorization
- **Symbolic Cholesky inversion**
  - determine actual non-zero structure of inverted Cholesky factor
  - determine supernodal structure of inverted Cholesky factor
  - determine storage space required for inversion
- **Numeric Cholesky inversion**
  - setup numerical data for inversion
  - perform numeric inversion

Figure 6. Brief summary of the algorithm for the computation of the Cholesky factorization and the inverted Cholesky factor.

would be forced to compute its Moore−Penrose inverse[68] in order to obtain numerically stable and meaningful results.

Fortunately, the Cholesky factorization is generally the most efficient method for testing positive definiteness of matrices,[69] and for overlap matrices, it might thus be best to try a regular Cholesky factorization as outlined further above. In case the Cholesky factorization fails, probably the best alternative in order to get around the demanding SVD[70] required for the computation of the Moore−Penrose inverse is to eliminate the linear dependencies among the basis functions. Although we have not yet implemented it, we want to note that this can be done by computing a LU[71] or QR[72] decomposition first and then using an iterative method for the determination of a basis of the null space.[73] Once this basis is known, it is possible to identify those columns that are linearly dependent and thus would cancel during the Cholesky factorization. These columns can then be permuted to the end of the matrix,[74] and the symbolic Cholesky factorization can then still be applied to the leading columns in order to reduce the fill-in.

Furthermore, we would like to note that for idempotent matrices the Cholesky factorization can also be computed via a QR factorization[75] and although we have not tested the reliability it has been found to yield acceptable results.[76]

7. Conclusion and Outlook

We have presented an efficient algorithm for the computation of the Cholesky factor and its corresponding inverse of a sparse symmetric positive definite matrix, the individual steps of which are briefly summarized in Figure 6. The high efficiency stems from splitting the factorization into a symbolic and a numeric part. The symbolic part is a strictly algebraic algorithm based on graph theory that first tries to find an appropriate reordering of the rows/columns in order to reduce the amount of fill-in and in a second step determines the exact nonzero structure of the resulting Cholesky factor. Furthermore this step finds blocks of columns which share a similar nonzero pattern, the so-called supernodes. These two steps are the most important features of this implementation and account for the superior efficiency of this algorithm as compared to a straightforward sparse implementation.

Furthermore, we have shown that this approach turns out to be useful for the computation of inverted Cholesky

factors as well, and we have stressed that it is crucial to use a ND reordering algorithm in order to preserve at least some degree of sparsity in the inverse. Although we have not made any attempt of implementation so far, we would like to stress that the ND reordering also allows for a parallel implementation.

Although the ND algorithm used for our algorithm is a rather small part of our code, we are convinced that this is the most important ingredient, and this divide and conquer strategy might also turn out to be useful for the design of other low-order scaling methods as well. For example, one might think of dividing a large molecule into many local domains based on decomposing two-center overlap matrices in the context of local correlation methods.

Acknowledgment. We thank Daniel S. Lambrecht and Eric J. Sundstrom for valuable discussions. This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Chemical Sciences Division of the U.S. Department of Energy under contract no. DE-AC0376SF00098. M.H.-G. is a part-owner of Q-CHEM Inc.

References

(1) Beebe, N. H. F.; Linderberg, J. Simplifications in the generation and transformation of two-electron integrals in molecular calculations. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.

(2) (a) Røeggen, I.; Wisløff-Nilssen, E. On the Beebe-Linderberg two-electron integral approximation. *Chem. Phys. Lett.* **1986**, *132*, 154–160. (b) O'Neal, D.; Simons, J. Application of Cholesky-like matrix decomposition methods to the evaluation of atomic orbital integrals and integral derivatives. *Int. J. Quantum Chem.* **1989**, *36*, 673–688. (c) Koch, H.; Sánchez de Merás, A.; Pedersen, T. B. Reduced scaling in electronic structure calculations using Cholesky decompositions. *J. Chem. Phys.* **2003**, *118*, 9481–9484. (d) Aquilante, F.; Pedersen, T. B.; Lindh, R. Low-cost evaluation of the exchange Fock matrix from Cholesky and density fitting representations of the electron repulsion integrals. *J. Chem. Phys.* **2007**, *126*, 194106. (e) Aquilante, F.; Malmqvist, P.-A.; Bondo Pedersen, T.; Ghosh, A.; Roos, B. O. Cholesky Decomposition-Based Multiconfiguration Second-Order Perturbation Theory (CD-CASPT2): Application to the Spin-State Energetics of Co^{III} (diiminato)(NPh). *J. Chem. Theory Comput.* **2008**, *4*, 694–702. (f) Aquilante, F.; Lindh, R.; Pedersen, T. B. Analytic derivatives for the Cholesky representation of the two-electron integrals. *J. Chem. Phys.* **2008**, *129*, 034106. (g) Røeggen, I.; Johansen, T. Cholesky decomposition of the two-electron integral matrix in electronic structure calculations. *J. Chem. Phys.* **2008**, *128*, 194107. (h) Weigend, F.; Kattannek, M.; Ahlrichs, R. Approximated electron repulsion integrals: Cholesky decomposition versus resolution of the identity methods. *J. Chem. Phys.* **2009**, *130*, 164106. (i) Chwee, T. S.; Carter, E. A. Cholesky decomposition within local multireference singles and doubles configuration interaction. *J. Chem. Phys.* **2010**, *132*, 074104.

(3) Millam, J. M.; Scuseria, G. E. Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic structure calculations. *J. Chem. Phys.* **1997**, *106*, 5569–5577.

(4) Schweizer, S.; Kussmann, J.; Doser, B.; Ochsenfeld, C. Linear-Scaling Cholesky decomposition. *J. Comput. Chem.* **2008**, *29*, 1004–1010.

(5) Aquilante, F.; Bondo Pedersen, T.; Sánchez de Merás, A.; Koch, H. Fast noniterative orbital localization for large molecules. *J. Chem. Phys.* **2006**, *125*, 174101.

(6) Zienau, J.; Clin, L.; Doser, B.; Ochsenfeld, C. Cholesky-decomposed densities in Laplace-based second-order Møller-Plesset perturbation theory. *J. Chem. Phys.* **2009**, *130*, 204112.

(7) Farkas, O.; Schlegel, B. H. Geometry optimization methods for modeling large molecules. *J. Mol. Struct. Theochem* **2003**, *666−667*, 31–39.

(8) Aquilante, F.; Pedersen, T. B. Quartic scaling evaluation of canonical scaled opposite spin second-order Møller Plesset correlation energy using Cholesky decompositions. *Chem. Phys. Lett.* **2007**, *449*, 354–357.

(9) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. Scaled opposite-spin second or-der Møller-Plesset correlation energy: An economical electronic structure method. *J. Chem. Phys.* **2004**, *121*, 9793–9802.

(10) (a) Aquilante, F.; Lindh, R.; Bondo Pedersen, T. Unbiased auxiliary basis sets for accurate two-electron integral approximations. *J. Chem. Phys.* **2007**, *127*, 114107. (b) Boman, L.; Koch, H.; Sánchez de Merás, A. Method specific Cholesky decomposition: Coulomb and exchange energies. *J. Chem. Phys.* **2008**, *129*, 134107. (c) Aquilante, F.; Gagliardi, L.; Pedersen, T. B.; Lindh, R. Atomic Cholesky decompositions: A route to unbiased auxiliary basis sets for density fitting approximation with tunable accuracy and efficiency. *J. Chem. Phys.* **2009**, *130*, 154107.

(11) Li, X.-P.; Nunes, R. W.; Vanderbilt, D. Density-matrix electronic-structure method with linear system-size scaling. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *47*, 10891–10894.

(12) (a) Challacombe, M. A simplified density matrix minimization for linear scaling self-consistent field theory. *J. Chem. Phys.* **1999**, *110*, 2332–2342. (b) Nunes, R. W.; Vanderbilt, D. Generalization of the density-matrix method to a nonorthogonal basis. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17611–17614. (c) Daniels, A. D.; Millam, J. M.; Scuseria, G. E. Semiempirical methods with conjugate gradient density matrix search to replace diagonalization for molecular systems containing thousands of atoms. *J. Chem. Phys.* **1997**, *107*, 425–431. (d) Bates, K. R.; Daniels, A. D.; Scuseria, G. E. Comparison of conjugate gradient density matrix search and Chebyshev expansion methods for avoiding diagonalization in large-scale electronic structure calculations. *J. Chem. Phys.* **1998**, *109*, 3308–3312. (e) Daniels, A. D.; Scuseria, G. E. What is the best alternative to diagonalization of the Hamiltonian in large scale semiempirical calculations? *J. Chem. Phys.* **1999**, *110*, 1321–1328. (f) Helgaker, T.; Larsen, H.; Olsen, J.; Jørgensen, P. Direct optimization of the AO density matrix in Hartree-Fock and Kohn-Sham theories. *Chem. Phys. Lett.* **2000**, *327*, 397–403. (g) Larsen, H.; Olsen, J.; Jørgensen, P.; Helgaker, T. Direct optimization of the atomic-orbital density matrix using the conjugate-gradient method with a multilevel preconditioner. *J. Chem. Phys.* **2001**, *115*, 9685–9697.

(13) (a) Ochsenfeld, C.; Head-Gordon, M. A reformulation of the coupled perturbed self-consistent field equations entirely within a local atomic orbital density matrix-based scheme. *Chem. Phys. Lett.* **1997**, *270*, 399–405. (b) Shao, Y.; Saravanan, C.; Head-Gordon, M. Curvy steps for density matrix-based energy minimization: Application to large-scale self-consistent-field calculations. *J. Chem. Phys.* **2003**, *118*, 6144–6151. (c) Head-Gordon, M.; Shao, Y.; Saravanan, C.; White, C. A. Curvy steps for density matrix based energy minimization: tensor formulation

and toy applications. *Mol. Phys.* **2003**, *101*, 37–43. (d) Ochsenfeld, C.; Kussmann, J.; Koziol, F. Ab Initio NMR Spectra for Molecular Systems with a Thousand and More Atoms: A Linear-Scaling Method. *Angew. Chem.* **2004**, *116*, 4585–4589. (e) Ochsenfeld, C.; Kussmann, J.; Koziol, F. Ab Initio NMR Spectra for Molecular Systems with a Thousand and More Atoms: A Linear-Scaling Method. *Angew. Chem., Int. Ed.* **2004**, *43*, 4485–4489.

(14) Guidon, M.; Hutter, J.; Vande Vondele, J. Auxiliary Density Matrix Methods for Hartree-Fock Exchange Calculations. *J. Chem. Theory Comput.* **2010**, *6*, 2348–2364.

(15) (a) Head-Gordon, M.; Maslen, P. E.; White, C. A. A tensor formulation of many-electron theory in a nonorthogonal single-particle basis. *J. Chem. Phys.* **1998**, *108*, 616–625. (b) Scuseria, G. E. Linear Scaling Density Functional Calculations with Gaussian Orbitals. *J. Phys. Chem. A* **1999**, *103*, 4782–4790.

(16) Jansík, B.; Høst, S.; Jørgensen, P.; Olsen, J.; Helgaker, T. Linear-scaling symmetric square-root decomposition of the overlap matrix. *J. Chem. Phys.* **2007**, *126*, 124104.

(17) Basic Linear Algebra Subprograms; http://www.netlib.org/blas. Accessed December 02, 2010.

(18) Linear Algebra Package; http://www.netlib.org/lapack. Accessed December 02, 2010.

(19) Shao, Y.; et al. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.

(20) George, A. In Algorithms for Large Scale Linear Algebraic Systems; *NATO ASI Series C: Mathematical and Physical Sciences*; Althaus, G. W., Spedicato, E., Eds.; Kluwer Academic Publishers: New York, 1998; Vol. 508; pp 73−105.

(21) (a) Chen, Y.; Davis, T. A.; Hager, W. W.; Rajamanickam, S. Algorithm 887: CHOLMOD, Supernodal Sparse Cholesky Factorization and Update/Downdate. *ACM T. Math. Software* **2008**, *35*, 22:1–14. (b) Davis, T. A.; Hager, W. W. Dynamic Supernodes in Sparse Cholesky Update/Downdate and Triangular Solves. *ACM T. Math. Software* **2009**, *35*, 27:1–23.

(22) (a) Schenk, O. *Scalable Parallel Sparse LU Factorization Methods on Shared Memory Multiprocessors*. Ph.D. thesis, ETH Zürich, 2000; (b) Schenk, O.; Gärtner, K. Solving unsymmetric sparse systems of linear equations with PARDISO. *Future Generat. Comput. Syst.* **2004**, *20*, 475–487. (c) Schenk, O.; Gärtner, K. On Fast Factorization Pivoting Methods for Sparse Symmetric Indefinite Systems. *Electron. T. Numer. Ana.* **2006**, *23*, 158–179.

(23) Liang, W.; Head-Gordon, M. An exact reformulation of the diagonalization step in electronic structure calculations as a set of second order nonlinear equations. *J. Chem. Phys.* **2004**, *120*, 10379–10384.

(24) Higham, N. J. Cholesky factorization. *WIREs Comput. Stat.* **2009**, *1*, 251–254.

(25) Higham, N. J. In *Reliable Numerical Computation*; Cox, M. G., Hammarling, S. J., Eds.; Oxford University Press: Oxford, U.K., 1990; pp 161−185.

(26) Gilbert, J. R. Predicting Structure in Sparse Matrix Computations. *SIAM J. Matrix Anal. Appl.* **1994**, *15*, 62–79.

(27) Higham, N. J.; Pothen, A. Stability of the partitioned inverse method for parallel solution of sparse triangular systems. *SIAM J. Sci. Comput.* **1994**, *15*, 139–148.

(28) Tinney, W.; Walker, J. Direct solutions of sparse network equations by optimally ordered triangular factorization. *Proc. IEEE* **1967**, *55*, 1801–1809.

(29) George, A.; Liu, J. W.-H. *Computer Solution of Large Sparse Positive Definite Systems*; Prentice-Hall: Englewood Cliffs, NJ, 1981.

(30) George, A.; Liu, J. W. H. An optimal algorithm for symbolic factorization of symmetric matrices. *SIAM J. Comput.* **1980**, *9*, 583–593.

(31) Ng, E.; Peyton, B. W. A supernodal Cholesky factorization algorithm for shared-memory multiprocessors. *SIAM J. Sci. Comput.* **1993**, *14*, 761–769.

(32) Liu, J. W. H.; Ng, E. G.; Peyton, B. W. On finding supernodes for sparse matrix computations. *SIAM J. Matrix Anal. Appl.* **1993**, *14*, 242–252.

(33) Liu, J. W. H. The role of elimination trees in sparse factorization. *SIAM J. Matrix Anal. Appl.* **1990**, *11*, 134–172.

(34) (a) Rose, D. J.; Tarjan, R. E.; Lueker, G. S. Algorithmic aspects of vertex elimination. *SIAM J. Comput.* **1976**, *5*, 266–283. (b) Tarjan, R. E.; Yannakakis, M. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.* **1984**, *13*, 566–579.

(35) Yannakakis, M. Computing the minimum fill-in is NP-complete. *SIAM J. Alg. Disc. Meth.* **1981**, *2*, 77–79.

(36) (a) Cuthill, E.; McKee, J. Reducing the bandwidth of sparse symmetric matrices. 1969;(b) Chan, W. M.; George, A. A linear time implementation of the reverse Cuthill-McKee algorithm. *BIT* **1980**, *20*, 8–14.

(37) Biermann, M. *Erkennen von Graphenklassen mittels lexikographischer Breitensuche*. M. Sc. Thesis, FernUniversität Hagen, 2007.

(38) (a) Liu, J. W. H. Modification of the minimum-degree algorithm by multiple elimination. *ACM T. Math. Software* **1985**, *11*, 141–153. (b) George, A.; Liu, W. H. The evolution of the minimum degree ordering algorithm. *SIAM Rev.* **1989**, *31*, 1–19.

(39) Amestoy, P. R.; Davis, T. A.; Duff, I. S. An Approximate Minimum Degree Ordering Algorithm. *SIAM J. Matrix Anal. Appl.* **1996**, *17*, 886–905.

(40) Khaira, M. S.; Miller, G. L.; Sheffler, T. J. Nested Dissection: A survey and comparison of various nested dissection algorithms; Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1992.

(41) (a) George, A. Nested Dissection of a regular finite element mesh. *SIAM J. Numer. Anal.* **1973**, *10*, 345–363. (b) Armon, D.; Reif, J. Space and time efficient implementations of parallel nested dissection. In *SPAA '92: Proceedings of the fourth annual ACM symposium on Parallel algorithms and architectures*, New York, NY, USA, 1992; pp 344−352.

(42) (a) Pothen, A.; Rothberg, E.; Simon, H.; Wang, L. Parallel sparse Cholesky factorization with spectral nested dissection ordering. In *Proceedings of the Fifth SIAM Conference on Applied Linear Algebra*, 1994; pp 418−422. (b) Schulze, J.; Diekmann, R.; Preis, R. Comparing nested dissection orderings for parallel sparse matrix factorization. In *Proceedings of PDPTA '95, CSREA 96-1103*, 1995; pp 280−289. (c) Bornstein, C. F.; Maggs, B. M.; Miller, G. L. Tradeoffs between parallelism and fill in nested dissection. In *SPAA '99: Proceedings of the eleventh annual ACM symposium on Parallel algorithms and architectures*, New York, NY, USA, 1999; pp 191−200. (d) Boman, E. G.; Wolf, M. M. A nested dissection approach to sparse matrix partitioning for parallel computations. Technical report, Sandia National Laboratories, NM, 2008.

(43) Karypis, G.; Kumar, V. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM J. Sci. Comput.* **1998**, *20*, 359–392.

(44) (a) Kussmann, J.; Ochsenfeld, C. Linear-scaling method for calculating nuclear magnetic resonance chemical shifts using gauge-including atomic orbitals within Hartree-Fock and density-functional theory. *J. Chem. Phys.* **2007**, *127*, 054103. (b) Kussmann, J.; Ochsenfeld, C. A density matrix-based method for the linear-scaling calculation of dynamic second- and third-order properties at the Hartree-Fock and Kohn-Sham density functional theory levels. *J. Chem. Phys.* **2007**, *127*, 204103.

(45) Heggernes, P.; Eisenstat, S. C.; Kumfert, G.; Pothen, A. The computational complexity of the minimum degree algorithm. In *Proceedings of 14th Norwegian Computer Science Conference, NIK 2001*, University of Troms, Norway. Also available as ICASE Report 2001-42, NASA/CR2001-211421, NASA Langley Research, pages 98–109.

(46) Conroy, J. M. Parallel nested dissection. *Parallel Comput.* **1990**, *16*, 139–156.

(47) Liu, J. W. A compact row storage scheme for Cholesky factors using elimination trees. *ACM T. Math. Software* **1986**, *12*, 127–148.

(48) Tarjan, R. E. Efficiency of a Good But Not Linear Set Union Algorithm. *J. ACM* **1975**, *22*, 215–225.

(49) Tarjan, R. E. Applications of Path Compression on Balanced Trees. *J. ACM* **1979**, *26*, 690–715.

(50) Tarjan, R. E. Data Structures and Network Algorithms. Number 44 in CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1983.

(51) Gilbert, J. R.; Ng, E. G.; Peyton, B. W. An Efficient Algorithm to Compute Row and Column Counts for Sparse Cholesky Factorization. *SIAM J. Matrix Anal. Appl.* **1994**, *15*, 1075–1091.

(52) Ackermann, W. Zum Hilbertschen Aufbau der reellen Zahlen. *Math. Ann.* **1928**, *99*, 118–133.

(53) Anderson, R. J.; Woll, H. Wait-free parallel algorithms for the union-find problem. In *STOC '91: Proceedings of the twenty-third annual ACM symposium on Theory of computing*, New York, NY, USA, 1991; pp 370–380.

(54) Tarjan, R. Depth-First Search and Linear Graph Algorithms. *SIAM J. Comput.* **1972**, *1*, 146–160.

(55) (a) Zmijewski, E.; Gilbert, J. A parallel algorithm for sparse symbolic Cholesky factorization on a multiprocessor. *Parallel Comput.* **1988**, *7*, 199–210. (b) Lewis, J. G.; Peyton, B. W.; Pothen, A. A fast algorithm for reordering sparse matrices for parallel factorization. *SIAM J. Sci. Stat. Comput.* **1989**, *10*, 1146–1173. (c) Geist, G. A.; Ng, E. Task scheduling for parallel sparse Cholesky factorization. *Int. J. Parallel Program.* **1990**, *18*, 291–314. (d) Gupta, A.; Kumar, V. A scalable parallel algorithm for sparse Cholesky factorization. In *Proceedings of the 1994 ACM/IEEE conference on Supercomputing*, Washington, D.C., 1994; pp 793–802. (e) Rothberg, E.; Schreiber, R. Improved load distribution in parallel sparse Cholesky factorization. In *Proceedings of the 1994 ACM/IEEE conference on Supercomputing*, Washington, D.C., 1994; pp 783–792. (f) Kumar, B.; Eswar, K.; Sadayappan, P.; Huang, C.-H. A reordering and mapping algorithm for parallel sparse Cholesky factorization. In *Proc. Scalable High Performance Computing Conference*, 1994.

(56) Liu, J. W. H. Reordering sparse matrices for parallel elimination. *Parallel Comput.* **1989**, *11*, 73–91.

(57) Liu, J. W. H. Equivalent sparse matrix reordering by elimination tree rotations. *SIAM J. Sci. Stat. Comput.* **1988**, *9*, 424–444.

(58) (a) Bird, R. S. On building trees with minimum height. *J. Funct. Program.* **1997**, *7*, 441–445. (b) Hsu, C.-H.; Peng, S.-L.; Shi, C.-H. Constructing a minimum height elimination tree of a tree in linear time. *Inf. Sci.* **2007**, *177*, 2473–2479.

(59) Manne, F. Reducing The Height Of An Elimination Tree Through Local Reorderings; Technical Report CS-51-91, University of Bergen, Norway, 1991.

(60) Pothen, A. The complexity of optimal elimination trees; Technical Report CS-88-16, Pennsylvania State University, USA, 1988.

(61) Benzi, M.; Tuma, M. Orderings for Factorized Sparse Approximate Inverse Preconditioners. *SIAM J. Sci. Comput.* **2000**, *21*, 1851–1868.

(62) Dongarra, J. In *Templates for the Solution of Algebraic Eigenvalue Problems: a Practical Guide*; Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H., Eds.; SIAM: Philadelphia, PA , 2000.

(63) van Grondelle, J. *Symbolic Sparse Cholesky Factorisation Using Elimination Trees*. M.Sc. Thesis, Utrecht University, 1999.

(64) Flake, J. Getting the Best from your Cache Architecture. *Info. Quarterly* **2004**, *3*, 14–15.

(65) Patterson, D. A.; Hennessy, J. *Computer Organization and Design*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 2004.

(66) (a) Goto, K.; van de Geijn, R. On reducing TLB misses in matrix multiplication. Technical Report TR-2002-55, University of Texas at Austin, USA, 2002. (b) Goto, K.; van de Geijn, R. A. Anatomy of a High-Performance Matrix Multiplication. *ACM T. Math. Software* **2008**, *34*, 1–25. (c) Goto, K.; van de Geijn, R. High Performance Implementation of the Level-3 BLAS. *ACM T. Math. Software* **2008**, *35*, 4:1–14.

(67) Scott, J.; Hu, Y.; Gould, N. In *Applied Parallel Computing*; Dongarra, J., Madsen, K., Wasniewski, J., Eds.; Springer: Berlin/Heidelberg, 2006; Vol. 3732 pp 818−827.

(68) (a) Moore, E. H. On the reciprocal of the general algebraic matrix. *B. Am. Math. Soc.* **1920**, *26*, 394–395. (b) Penrose, R. A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.* **1955**, *51*, 406–413.

(69) (a) Hansen, P. C. Detection of near-singularity in Cholesky and LDLT factorizations. *J. Comput. Appl. Math.* **1987**, *19*, 293–299. (b) Barlow, J. L.; Vemulapati, U. B. Rank Detection Methods for Sparse Matrices. *SIAM J. Matrix Anal. A.* **1992**, *13*, 1279–1297.

(70) (a) Golub, G.; Kahan, W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *J. Soc. Ind. Appl. Math. B* **1965**, *2*, 205–224. (b) Loan, C. F. V. Generalizing the Singular Value Decomposition. *SIAM J. Numer. Anal.* **1976**, *13*, 76–83. (c) Golub, G.; Sølna, K.; Dooren, P. V. Computing the SVD of a General Matrix Product/Quotient. *SIAM J. Matrix Anal. A* **2000**, *22*, 1–19.

(71) Gilbert, J. R.; Ng, E. G. Predicting structure in nonsymmetric sparse matrix factorizations. In *Graph Theory and Sparse Matrix Computation*; Springer-Verlag, 1992; pp 107–139.

(72) Davis, T. Multifrontral multithreaded rank-revealing sparse QR factorization. In *Combinatorial Scientific Computing*, number 09061; Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2009. Naumann, U., Schenk, O., Simon, H. D.,

Toledo, S., Eds.; Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany.

(73) (a) Berry, M. W.; Heath, M. T.; Kaneko, I.; Lawo, M.; Plemmons, R. J.; Ward, R. C. An Algorithm to Compute a Sparse Basis of the Null Space. *Numer. Math.* **1985**, *47*, 483–504. (b) Foster, L. V. Rank and Null Space Calculations Using Matrix Decompositions without Column Interchanges. *Linear Algebra Appl.* **1986**, *74*, 47–71. (c) Choi, S.-C. *Iterative Methods for Singular Linear Equations and Least-Squares Problems*, Ph.D. Thesis, Stanford University, CA, 2006; (d) Le Borne, S. Block computation and representation of a sparse nullspace basis of a rectangular matrix. *Linear Algebra Appl.* **2008**, *428*, 2455–2467. (e) Gotsman, C.; Toledo, S. On the Computation of Null Spaces of Sparse Rectangular Matrices. *SIAM J. Matrix Anal. A* **2008**, *30*, 445–463. (f) Wu, J.; Lee, Y.-J.; Xu, J.; Zikatanov, L. Convergence Analysis on Iterative Methods for Semidefinite Systems. *J. Comput. Math.* **2008**, *26*, 797–815.

(74) Arbenz, P.; Drmac, Z. On Positive Semidefinite Matrices with Known Null Space. *SIAM J. Matrix Anal. Appl.* **2002**, *24*, 132–149.

(75) Moler, C. B.; Stewart, G. W. On the Householder-Fox algorithm for decomposing a projection. *J. Comput. Phys.* **1978**, *28*, 82–91.

(76) Fox, K.; Krohn, B. J. Computation of cubic harmonics. *J. Comput. Phys.* **1977**, *25*, 386–408.

CT100618S

JCTC Journal of Chemical Theory and Computation

Bond Length Alternation of Conjugated Oligomers: Wave Function and DFT Benchmarks

Denis Jacquemin*,† and Carlo Adamo*,‡

Laboratoire CEISAM − UMR CNR 6230, Université de Nantes, 2 Rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France, and Ecole Nationale Supérieure de Chimie de Paris, Laboratoire Electrochimie et Chimie Analytique, UMR CNRS-ENSCP no. 7575, 11, rue Pierre et Marie Curie, F-75321 Paris Cedex 05, France

Received November 12, 2010

Abstract: We have computed the bond length alternation (BLA) in a series of π-conjugated quasilinear chains containing from two to six unit cells. Several structures (eight oligomeric sets including three conformers of polyacetylene, polymethineimine, polysilaacetylene, etc.) have been considered to cover the possible evolutions of the BLA with increasing chain length. Three objectives have been tackled: (1) the computation of accurate reference values using the CCSD(T) theory; (2) an evaluation of the performances of other electron correlated wave function approaches (MP*n*, SCS-MP2, CCSD, etc.); (3) the benchmarking of several DFT functionals, including global, range-separated, and double hybrids. It turns out that the SCS-MP2 approach is, on average, an efficient scheme in terms of its accuracy/cost ratio. Among the selected DFT approaches, no single functional emerges as uniformly accurate for all oligomeric series and chain lengths, but BHHLYP, M06-2X, and CAM-B3LYP could be reasonable choices for long oligomers.

1. Introduction

The bond length alternation (BLA) is a geometrical parameter calculated as the difference between the lengths of a single bond and the adjacent multiple (double or triple) bond in π-delocalized systems. For polyacetylene (PA, referred to as CC-II in the following), a polymer constituted of a sequence of sp^2 carbon atoms (see Figure 1), it is well-known that there is a close connection between the BLA and the electronic gap.[1,2] Indeed, in the one-electron approximation, there is a simple proportionality relationship between these two properties.[3] More generally, in π-conjugated compounds, the geometric and electronic structures are closely related, and an accurate description of the ground-state structures is an actual prerequisite for the determination of valid electronic properties.[4,5] The interested reader may find examples of the key role played by the BLA in several domains, including nonlinear optics,[6−8] two-photon absorption efficiencies,[9−11] transport properties,[12] and photochromic features.[13,14]

Straightforward experimental determinations of the BLA remain difficult, and the results may be relatively disappointing. First, gas-phase measurements are only possible on the shortest oligomers, as the intermolecular interaction energies tend to be substantial in π systems. This is unsatisfying, as the BLA evolves slowly with chain length; e.g., it is roughly divided by two when going from butadiene to infinitely long polyacetylene.[3,15] In other words, chain end effects cannot be neglected, as they span over more than five unit cells (*N*).[16] Second, the available XRD data[17] suffer from the impact of environmental effects that are sizable for π-rich oligomers.[18] This statement can also be illustrated by comparisons between gas-phase and condensed-phase simulations that yield strongly dissimilar BLA in some cases.[19] Eventually, the experimental accuracy can also be a limiting parameter; e.g., two gas-phase experiments carried out on the simple and symmetric *trans*-butadiene yielded BLAs (double/single bond distances) of 0.118 Å (1.349/1.467 Å)[20]

* E-mail: Denis.Jacquemin@univ-nantes.fr (D. J.); carlo-adamo@enscp.fr (C. A.).

† CEISAM, Nantes.

‡ ENSCP, Paris.

10.1021/ct1006532
Published on Web 12/29/2010

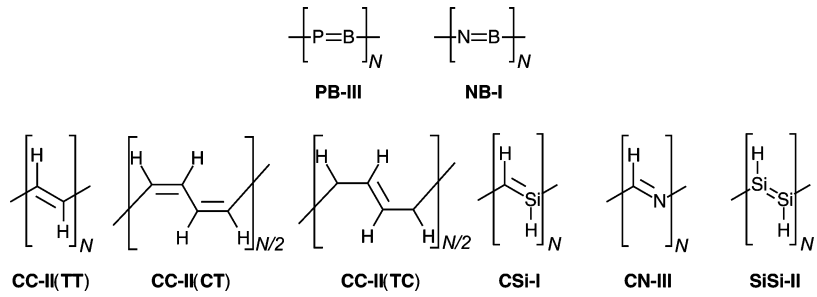


Figure 1. Representation of the oligomers considered in this work. All chains are capped by terminal hydrogen atoms. *N* is the number of unit cells.

and 0.130 Å (1.337/1.467 Å),[21] a 10% discrepancy. The same holds for solid-sate measurements obtained through XRD[22] or NMR[23] techniques that allow for getting useful, yet not very accurate, estimates of the BLA of the polymeric PA: 0.08 ± 0.03 Å.[18]

Due to these limitations, numerous efforts have been devoted to the computation of the BLA of conjugated chains with reliable theoretical models. For PA chains, the most intensively studied quasilinear oligomers, the first systematic study dates from 1997[3] and demonstrated that Hartree−Fock (HF) predicts BLA too large by a factor of 2, whereas a pure Density Functional Theory (DFT) scheme, namely BLYP,[24,25] leads the opposite error. On the contrary, the geometrical parameters obtained by the second-order Møller−Plesset (MP2) and the hybrid B3LYP functional[26] have been found in satisfying agreement with experimental measurements on short chains.[3] These conclusions were later confirmed by, on the one hand, simulations relying on more recent global hybrid functionals[27] and, on the other hand, self-interaction corrected DFT schemes.[28] However, during the past three years, it has been shown that CCSD(T) and MP2 BLAs are far from coinciding when the polyene chain lengthens.[29−31] Additionally, different authors have indicated that range-separated hybrids[30,32,33] as well as the spin-component scaled MP2 (SCS-MP2) approach[31] are in fact more accurate than conventional hybrids like B3LYP for PA. The difficulties encountered when modeling the BLA of polymethineimine (PMI, referred to as CN-III in Figure 1) are much more dramatic, though this polymer is isoelectronic to PA. Indeed, for PMI, global hybrids and MP2 already provide diverging results,[34] and the DFT estimates appear to decrease too rapidly with the chain length. This finding also holds, but to a smaller extent, for range-separated hybrids[30] and self-interaction corrected schemes:[35] no available DFT approach is able to completely cure the too sharp falloff, though the most recent functionals, combining range-separation and second-order perturbative corrections,[36] clearly attenuate the problem.[33]

Five years ago, we investigated several series of oligomers with MP2[37] and DFT[38] approaches. These works allowed for the definition of three phenomenological categories. In the first set (type I oligomers), the BLA exponentially decreases with the chain length and rapidly converges to zero. In the second category (type II), one finds symmetric oligomers that, due to the Peierls distortion, show nonzero BLA for all chain lengths. The last class (type III) is constituted of asymmetric compounds presenting significantly different bond lengths even in very long oligomers. For type I, most *ab initio* models, including conventional hybrid functionals, give relatively accurate values,[37,38] whereas for type II, the *exact* exchange balance seems essential, as illustrated by the above discussion for PA. Eventually, for type III oligomers, no classic DFT functionals seem completely satisfying (see PMI above).[38] However, these previous works have been relying on MP2 references values[38] that are far from flawless. Additionally, other investigations performed with more refined wave function or DFT schemes have been limited to PA and PMI,[3,27,28,30−35] making any general conclusions difficult, if not impossible. In this paper, we treat several series (see Figure 1) of oligomers and go significantly beyond previous studies by (1) computing CCSD(T) BLA for small and medium oligomers of type I (NB-I and CSI-I), II (CC-II and SiSi-II), and type III (PB-III and CN-III) oligomeric series; (2) evaluating the performances of several electron correlated wave function schemes, including CCSD and SCS-MP2 approaches; and (3) assessing the efficiency of DFT functionals (pure as well as global, range-separated, and double hybrids).

2. Method

All calculations have been performed with the Gaussian 09 program,[39] except for the SCS-MP2[40] and B2PLYP[41−43] calculations, which have been achieved with the ORCA code.[44] We have systematically used a tightened SCF threshold (10^{-10} au) and geometry optimization criteria (rms force smaller than 10^{-5} au). The 6-31G(d) basis set has been selected throughout (see next section). HF, MP2, MP4(SDQ), CCSD, MP4, and CCSD(T) calculations have been performed using analytic gradients, except for the two latter approaches, which relied on numerical differentiation. Consequently, the MP4 and CCSD(T) calculations have been the clear time-limiting steps in the present investigation (more than one year of CPU time for the hexamer of CN-III at the MP4 level). Several DFT functionals have also been used. First, we compared methods presenting a constant correlation functional (LYP): one GGA, BLYP;[24,25] two global hybrids, namely, B3LYP[26] and BHHLYP;[45] two range-separated hybrids, LC-BLYP[46] and CAM-B3LYP;[47] as well as a double-hybrid, that is, B2PLYP.[41−43] For the record, note that the original damping parameter of 0.33 au for LC-BLYP has been applied to allow consistent comparisons with our previous work.[30] Second, we also considered four extra modern functionals: B97-D,[48] which is free of *exact ex-*

change; two members of the M06 family (M06 and M06-2X);[49] as well as a recent range-separated hybrid (ωB97).[50] The optimizations have been performed by taking the symmetry into account but fully optimizing all nonredundant distances and valence angles (including the hydrogen-related bonds and angles). The only exceptions are CSI-I and CN-III, for which the valence angles of the skeleton have been set equal in order to enforce quasi-linearity of the molecules. Such a scheme has already been applied for the same two systems, and we refer the reader to these previous works for discussion.[30,35,37,38] For SCS-MP2 and B2PLYP, the skeleton valence angles have been fixed to their MP2 values for both CSI-I and CN-III. Test calculations have shown that this approximation has a negligible impact on the BLA. The BLAs reported in the following have been systematically measured at the center of the oligomers, as they represent a better approximation of the behavior obtained in longer chains. Note that all chains of Figure 1 are capped by terminal hydrogen atoms during our simulations.

3. Results

3.1. Reference CCSD(T) Values. Before analyzing the results obtained for different oligomers, it is worth it to discuss the choice of the 6-31G(d) basis set. This selection of a relatively compact basis set is dictated by our consideration of the hexamer with numerical MP4 and CCSD(T) derivatives, but is clearly sound in view of several previous investigations on the topic. Indeed, for all-*trans* CC-II, the difference between the MP2/6-31G(d) and MP2/cc-pVDZ polymeric BLA is as small as 0.002 Å,[18] whereas for the octamer, the discrepancy between the MP2/6-31G(d) and MP2/6-311G(3df) BLA is only 0.003 Å.[37] For the dimer and tetramer of CC-II, the CCSD(T) differences between the 6-31G(d) and "best estimates" are both 0.003 Å according to the evaluations of Zhao and Truhlar.[29] For the CN-III dimer, the CCSD(T) BLA evolves only by +0.004 Å, when going from 6-31G(d) to 6-311G(3df).[35] It is also noticeable that, for the same system, the corresponding B3LYP (+0.005 Å), MP2 (+0.003 Å), and MP4 (+0.004 Å) basis set shifts are completely similar.[35] In addition, for the same system, the difference between the CCSD and CCSD(T) BLA systematically amounts to +0.007 Å with 6-31G(d), 6-311G(2d), and 6-311G(3df). For the trimer of PMI, the CCSD/CCSD(T) BLA difference is also nearly constant with 6-31G(d) at +0.007 Å and 6-311G(2d) at +0.008 Å. These remarkably stable results hint that the quite large errors of CCSD for type III chains (see below) are not basis-set related. In other words, while using much larger basis sets would imply small variations of the computed CCSD(T) values (ca. 3×10^{-3} Å), such a choice would not affect the conclusions regarding the relative accuracies of the different computational schemes nor the chemical trends noted below. The interested reader may find a detailed basis set study at the MP2 level in ref 37, as well as a complete coupled cluster investigation (including large Dunning's basis set) for butadiene in ref 31, and both works allow for the concludsion that 6-31G(d) is indeed a very good compromise for the BLA.

Table 1. CCSD(T)/6-31G(d) BLA (Å) Computed for the Oligomers Sketched in Figure 1[a]

system	$N = 2$	$N = 3$	$N = 4$	$N = 5$	$N = 6$
PB-III	0.1493	0.1272	0.1164	0.1103	0.1053
NB-I	0.1414	0.1179	0.0988	0.0885	0.0764
CC-II (TT)	0.1154	0.1022	0.0941	0.0905	0.0880
CC-II (CT)	0.1154	0.1017	0.0951	0.0912	0.0893
CC-II (TC)	0.1284	0.1151	0.1060	0.1018	0.0992
CSI-I	0.1016	0.0951	0.0848	0.0786	0.0723
CN-III	0.1356	0.1282	0.1181	0.1138	0.1097
SiSi-II	0.1092	0.0948	0.0888	0.0855	0.0836

[a] See Figure 2 for a graphical representation of these data.

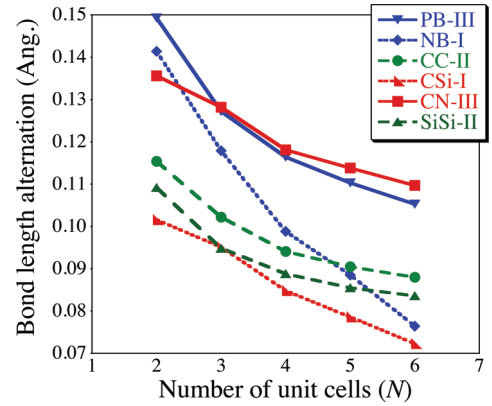


Figure 2. Evolution from the dimer to the hexamer of the CCSD(T) BLA. For CC-II, only the all-trans (TT) conformation has been represented.

The computed CCSD(T) values are collated in Table 1, and their evolutions with chain length are displayed in Figure 2. As expected, the BLA systematically decreases when the chain lengthens, as a result of the improved delocalization of the π electrons. Although the chains considered in this work are too short to allow direct extrapolation to the infinite oligomer limit,[16] one clearly sees that the BLA of NB-I and CSI-I rapidly declines, which is consistent with a type I evolution. Indeed, for the hexamer, they are the only two compounds with a central BLA smaller than 0.08 Å (see Table 1). This therefore qualitatively confirms a previous MP2 analysis.[27] From Figure 2, one clearly notes that the two type III oligomers, PB-III and CN-III, have significantly larger BLAs than the two type II derivatives, CC-II and SiSi-II. Nevertheless, the evolution of the BLA between the dimer and the hexamer remains similar for PA (−0.0274 Å) and PMI (−0.0259 Å). CC-II and SiSi-II also present BLA evolving at the same rate (−0.0256 Å for the latter), though their BLA tend to become slightly more alike as the chain lengthens. Eventually, from Table 1, one notes that going from the all-trans (TT) to the cis−trans (CT) conformer has a negligible impact on the BLA (ca. 1×10^{-3} Å), whereas the trans−cis (TC) oligomer possesses more dissimilar double and single bonds (BLA larger by ca. 1×10^{-2} Å).

3.2. Wave Function Benchmarks. The BLA computed with the HF, MP2, SCS-MP2, MP4(SDQ), MP4, and CCSD approaches can be found in the Supporting Information (SI), whereas the data obtained through a statistical analysis are given in Table 2. This table lists the mean signed errors (MSE) and mean absolute errors (MAE) obtained for different subsets, considering CCSD(T) as a reference. For

Table 2. Statistical Analysis of the Wavefunction Results[a]

	mean signed error					
set	HF	MP2	SCS-MP2	MP4(SDQ)	MP4	CCSD
$N = 2$	−293	4	−48	−79	21	−91
$N = 4$	−285	49	−39	−109	45	−122
$N = 6$	−271	80	−29	−97	61	−131
type I	−34	−3	−34	−51	−13	−43
type II	−385	56	−40	−115	28	−136
type III	−330	71	−43	−119	130	−153
full	−283	45	−39	−100	43	−117

	mean absolute error					
set	HF	MP2	SCS-MP2	MP4(SDQ)	MP4	CCSD
$N = 2$	293	30	48	79	34	91
$N = 4$	292	54	39	109	57	122
$N = 6$	313	80	29	111	68	131
type I	88	14	34	51	25	43
type II	385	66	40	121	31	136
type III	330	71	43	119	130	153
full	297	54	39	103	54	117

[a] CCSD(T) values have been used as references. Mean signed error [CCSD(T)-tested method] and mean absolute errors are given in 10^{-4} Å. Type I corresponds to NB-I and CSI-I, type II to CC-II (three conformers) and SiSi-II, and type III to PB-III and CN-III. See text for more details.

the full set of 40 derivatives, it is clear that HF significantly overshoots the BLA. This behavior was expected, as it is well-known that HF tends to provide a too localized picture for most organic compounds.[3,27] The typical absolute errors brought by HF are on the order of 3×10^{-2} Å, and the only series for which HF could be viewed as a reasonable approximation corresponds to the easiest case, that is, type I oligomers. In short, HF is inadequate and should not be used to compute the BLA. Among electron-correlated approaches, the most accurate scheme is SCS-MP2, closely followed by MP2 and MP4, whereas MP4(SDQ) and CCSD produce significantly larger errors. This clearly indicates the importance of a balance between double, triple and quadruple contributions for the most refined approaches. Interestingly, MP4(SDQ) and CCSD both overshoot the BLA. Although this phenomena is less pronounced than for HF, the average errors (1×10^{-2} Å) remain incompatible with accurate estimates.

The evolution with chain length of the errors can be appreciated by comparing the average errors obtained for dimers ($N = 2$), tetramers ($N = 4$), and hexamers ($N = 6$). It turns out that the discrepancies tend to slightly increase with chain length for all methods but SCS-MP2. For instance, the MAE doubles (almost triples) for MP4 (MP2) when going from $N = 2$ to $N = 6$. This explains why the MP2 scheme was previously considered very accurate when CCSD(T) calculations were only technically possible for the shortest chains. Comparing the MSE and MAE obtained for different categories of compounds is also enlightening: the deviations obtained for type I are systematically small; any theoretical scheme seems satisfying. Additionally, for this series, the deviations tend to decrease with the chain length, as all approaches predict a zero BLA when $N \rightarrow \infty$. The errors obtained for types II and III are larger than for type I and are similar for the two series, but MP4 is extremely efficient for the former but apparently misses the target for III. This phenomenon is associated with only one system, namely PB-III, as the MP4 errors for CN-III are very small (MSE and MAE of 0.0019 Å).

To obtain further insight, we have plotted in Figure 3 the evolutions with chain length of the errors for four selected derivatives, including examples of different types. For CC-II, extensive comparisons between MP2, SCS-MP2, and CCSD(T) values may also be found in ref 31. In all cases, the MP4(SDQ) and CCSD curves are almost coinciding. It is certainly striking that, for both CC-II and SiSi-II, two systems subject to Peierls distortion, the error patterns are similar and no wave function scheme possesses a flat curve; i.e., none of the tested approaches provides a constant error when the chain lengthens. Of course, in very long chains, this should become the case, but even for the hexamer, no convergence pattern clearly emerges for CC-II and SiSi-II. Nevertheless, one can predict, from the two top panels of Figure 3, that in long oligomers, SCS-MP2 is certainly more adequate than MP2 (and probably also than MP4). This finding backs the conclusions of Sancho-Garcia and Perez-Jimenez.[31] However, for infinitely long CC-II, SCS-MP2 probably slightly overestimates BLA, contrary to what is found in small polyenes. For CN-III, the MP2 and MP4 schemes provide nearly stabilized errors for $N = 4$, 5, and 6. If SCS-MP2 outperforms MP2 for short chains, the situation is hardly foreseeable for larger N. For CSI-I, all patterns are similar, the discrepancies remaining relatively small, as expected. Comparing the data in Table 2 to that in Figure 3 demonstrates that the nearly constant MSE and MAE with N noticed for SCS-MP2 are (in part) due to a compensation of errors between different series of oligomers. Eventually, the impact of changing the conformation is correctly predicted by all approaches; e.g., for the hexamer of CC-II, going from TT to CT (TC) induces a BLA increase of 0.0013 Å (0.0112 Å) at the CCSD(T) level, and MP2, SCS-MP2, MP4, and CCSD respectively deliver 0.0016 Å (0.0124 Å), 0.0010 Å (0.0137 Å), 0.0017 Å (0.0123 Å), and 0.0010 Å (0.0126 Å).

Overall, it seems that SCS-MP2 is an excellent compromise in terms of the accuracy/efficiency ratio, as it yields relatively small discrepancies that are quite uniformly distributed among the tested series and sizes. In fact, the MAE obtained with SCS-MP2 is on the order of the expected basis set error for the CCSD(T) reference values (see above). Nevertheless, SCS-MP2 tends to produce slightly too large of a BLA, and the error curves of Figure 3 are not flat. MP4 yields values close to the CCSD(T) reference, but in one specific case (PB-III). However, the computational effort associated with full MP4 (including contribution from the triples) is largely exceeding its SCS-MP2 counterpart. Indeed, the MP4 calculations involve resources similar to their CCSD(T) counterparts.

3.3. Density Functional Theory Benchmarks. As for the wave function results, the BLA computed with the selected six DFT functionals are given in the SI, whereas Table 3 collates the corresponding MSE and MAE. We have used a panel functional relying on the same correlation functional (LYP), to allow investigations of the impact of the exchange form, but we have additionally considered four

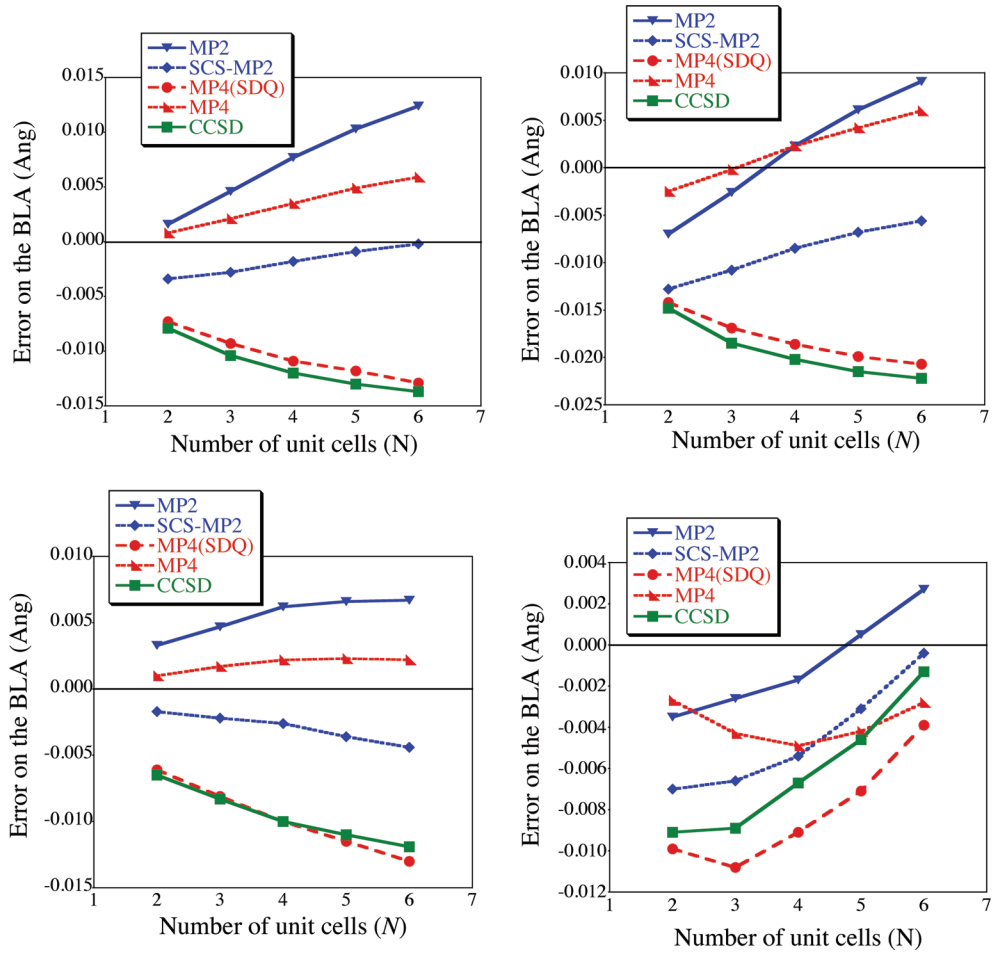


Figure 3. Evolution with the chain length of the errors obtained with wave function approaches for CC-II (top left), SiSi-II (top right), CN-III (bottom left), and CSI-I (bottom right). Note the different scales.

Table 3. Statistical Analysis of the DFT Results[a]

	mean signed error									
	BLYP	B3LYP	BHHLYP	LC-BLYP	CAM-B3LYP	B2PLYP	B97-D	M06	M06-2X	ωB97D
$N = 2$	102	11	−104	−151	−110	14	104	1	−96	−204
$N = 4$	259	116	−51	−146	−70	94	261	87	−58	−198
$N = 6$	337	170	−20	−142	−55	135	347	131	−38	−185
type I	213	117	43	−10	18	77	201	103	21	−20
type II	221	71	−116	−216	−134	60	231	58	−111	−277
type III	302	153	−35	−145	−54	137	301	118	−43	−214
full	239	103	−56	−147	−76	83	241	84	−61	−197

	mean absolute error									
	BLYP	B3LYP	BHHLYP	LC-BLYP	CAM-B3LYP	B2PLYP	B97-D	M06	M06-2X	ωB97D
$N = 2$	102	47	111	154	118	39	104	30	97	205
$N = 4$	259	116	76	155	86	94	261	87	72	204
$N = 6$	337	170	76	169	81	135	347	131	76	221
type I	213	117	70	55	55	77	200	108	49	67
type II	221	85	116	216	134	69	231	66	111	277
type III	302	153	37	145	54	137	301	119	43	214
full	239	110	85	158	94	88	241	90	79	209

[a] See caption of Table 2 for more details.

recently designed functionals (B97-D, M06, M06-2X, and ωB97). BLYP and B97-D yield very similar figures, and both systematically underestimate the BLA. The discrepancies with respect to CCSD(T) rapidly increase with the chain length; e.g., they reach a factor of 2 for the hexamer of CSI-I. Therefore, this work confirms that pure functionals are inadequate for estimating the geometry of medium and large π-conjugated chains. B3LYP and B2PLYP, respectively the most popular global and double hybrids, show similar patterns, an outcome already noticed for PA.[31] M06 follows alike trends but with errors systematically smaller than those of B3LYP. These three functionals predict BLAs that are

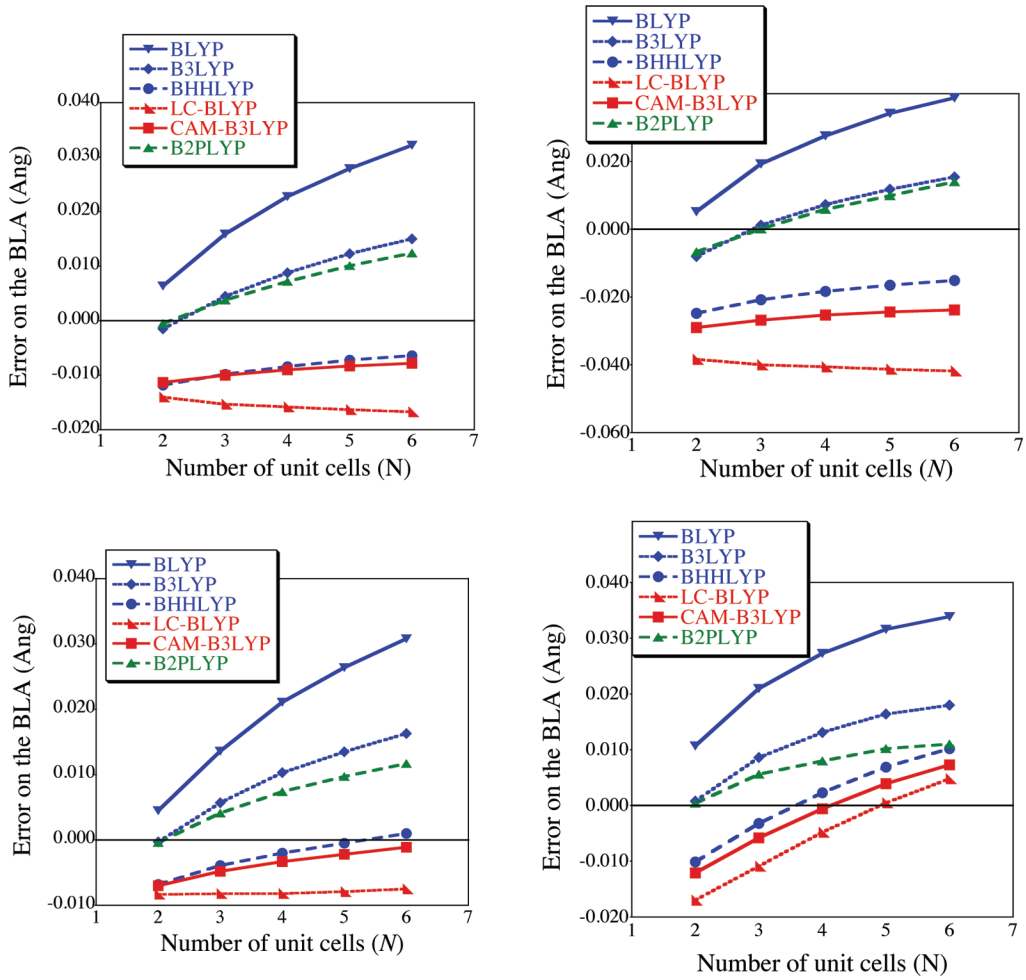


Figure 4. Evolution with chain length of the errors obtained with several DFT functionals for CC-II (top left), SiSi-II (top right), CN-III (bottom left), and CSI-I (bottom right). Note the different scales.

too small, with very small discrepancies for the dimers (for which they are the most accurate functionals, M06 providing the same performance as MP2), but their predictions tend to become less reliable with increasing chain length. For type II chains, they yield BLAs closer to the spot, but the errors nevertheless grow with N; i.e., they cannot be considered completely satisfying even for symmetric polymers. LC-BLYP and ωB97 enjoy relatively constant deviations with the chain length but significantly overshoot the BLA, especially the latter. Though the extent of this error is smaller than with HF (see previous section), the average absolute deviation still exceeds the 1×10^{-3} Å limit. Eventually, BHHLYP, CAM-B3LYP, and M06-2X deliver a more balanced description of the single and double bond lengths with stable MAEs for the tetramer and hexamer and good performances for the difficult case of type III oligomers. Overall, M06-2X is the most efficient functional, though as BHHLYP and CAM-B3LYP, it is not completely satisfying for type II chains, especially for SiSi-II. It is worth mentioning that CAM-B3LYP was already pinpointed as one of the most efficient functionals for the two prototype systems, PA[30,32] and PMI.[30]

In Figure 4, evolutions of the DFT (LYP correlation) error patterns are shown for four typical systems. The error ranges are significantly larger than in the corresponding Figure 3, and one notes that for all functionals and systems, the errors tend to increase with N, so that the conclusions obtained with short oligomers do not pertain to longer chains. The only noticeable exception is LC-BLYP, which allows for relatively constant errors in three out of four cases: CC-II, CN-III, and SiSi-II. From Figure 4, it seems obvious that a large share of exact exchange is needed to grant reasonable estimates of the infinite chain limit, though no functional can be considered completely satisfying. Therefore, the use of BHHLYP, M06-2X or CAM-B3LYP, which yield very similar values for most oligomers, may be a good choice for medium and large conjugated chains in spite of the large errors obtained for the dimers. It is noticeable that, in the long range, the three functionals present a similar share of exact exchange (50% for BHHLYP, 56% for M06-2X, and 65% for CAM-B3LYP). This investigation illustrates how a simple structural parameter can be difficult to predict even with refined DFT schemes.

As noted above, the conformational CCSD(T) increase in the BLA noted for the hexamer of CC-II is 0.0013 Å (0.0112 Å) for CT (TC). The corresponding values are 0.0023 Å (0.0066 Å), 0.0024 Å (0.0094 Å), 0.0020 Å (0.0116 Å), 0.0007 Å (0.0107 Å), 0.0013 Å (0.0111 Å), and 0.0022 Å (0.0108 Å) for BLYP, B3LYP, BHHLYP, LC-BLYP, CAM-B3LYP, and B2PLYP, respectively. B97-D, M06, M06-2X, and ωB97 respectively deliver 0.0034 Å (0.0054 Å), 0.0029 Å (0.0098 Å), 0.0017 Å (0.0118 Å), and 0.0011 Å (0.0111

Å), respectively. Therefore, BLYP and B97-D are again off target, whereas all other functionals give relatively accurate estimates of the impact of conformation.

4. Conclusions

We have computed the central bond length alternation in eight oligomeric series of increasing size. Reference CCSD(T) values have been obtained in each case, allowing accurate comparisons for several chain lengths and atomic compositions of the unit cell. It turns out that the error patterns of the tested wave function and density functional approaches are significantly affected by the considered system. Nevertheless, several general trends have been identified. As expected, HF produces BLA that are too large with absolute deviations of ca. 3×10^{-2} Å. Both MP4(SDQ) and CCSD also overshoot the BLA, and though the errors are smaller than for HF, they remain sizable (ca. 1×10^{-2} Å). MP2, MP4, and SCS-MP2 generally yield accurate BLAs (average absolute deviations of ca. 5×10^{-3} Å), the two former (the latter) slightly underestimating (overestimating) the reference data. SCS-MP2 is, on average, the most efficient, although the discrepancies with respect to the CCSD(T) value evolve with the chain length, at least up to the hexamer. None of the selected functionals has a net edge for the full set of molecules. Indeed, for long oligomers, it is obvious that a global hybrid including a large share to exact exchange (BHHLYP or M06-2X) or a balanced range-separated hybrid (CAM-B3LYP) could be smart default choice to evaluate the BLA, whereas, for the smallest chains, B3LYP or B2PLYP are to be preferred.

Acknowledgment. D.J. is indebted to the Région des Pays de la Loire for financial support in the framework of a recrutement sur poste stratégique. This research used resources of the Interuniversity Scientific Computing Facility located at the University of Namur, Belgium, which is supported by the F.R.S.-FNRS under convention No. 2.4617.07.

Supporting Information Available: Tables listing the computed BLA for all oligomers and methods. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Kertesz, M. *Chem. Phys.* **1979**, *44*, 349–356.

(2) Brédas, J. L. *J. Chem. Phys.* **1985**, *82*, 3808–3811.

(3) Choi, C. H.; Kertesz, M.; Karpfen, A. *J. Chem. Phys.* **1997**, *107*, 6712–6721.

(4) Brédas, J. L. *Adv. Mater.* **1995**, *7*, 263–274.

(5) Roncali, J. *Macromol. Rapid Commun.* **2007**, *28*, 1761–1775.

(6) Kirtman, B.; Champagne, B.; Bishop, D. M. *J. Am. Chem. Soc.* **2000**, *122*, 8007–8012.

(7) Jacquemin, D.; Perpete, E. A.; André, J. M. *J. Chem. Phys.* **2004**, *120*, 10317–10327.

(8) Murugan, N.; Kongsted, J.; Rinkevicius, Z.; Agren, H. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 16453–16458.

(9) Bartkowiak, W.; Zalesny, R.; Leszczynski, *J. Chem. Phys.* **2003**, *287*, 103–112.

(10) Zhang, G. P.; Sun, X.; George, T. F. *J. Phys. Chem. A* **2009**, *113*, 1175–1188.

(11) Reeve, J. E.; Anderson, H. L.; Clays, K. *Phys. Chem. Chem. Phys.* **2010**, *12*, 13484–13498.

(12) Wen, S. H.; Deng, W. Q.; Han, K. L. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9267–9275.

(13) Perrier, A.; Maurel, F.; Aubard, J. *J. Photochem. Photobiol. A: Chem.* **2007**, *189*, 167–176.

(14) Patel, P. D.; Masunov, A. E. *J. Phys. Chem. A* **2009**, *113*, 8409–8414.

(15) Jacquemin, D.; Perpete, E. A.; Ciofini, I.; Adamo, C. *Chem. Phys. Lett.* **2005**, *405*, 376–381.

(16) Schmalz, T. G.; Griffin, L. L. *J. Chem. Phys.* **2009**, *131*, 224301.

(17) Zhu, Q.; Fischer, J. E.; Zuzok, R.; Roth, S. *Solid State Commun.* **1992**, *83*, 179–183.

(18) Pino, R.; Scuseria, G. *J. Chem. Phys.* **2004**, *121*, 8113–8119.

(19) Cammi, R.; Mennucci, B.; Tomasi, J. *J. Am. Chem. Soc.* **1998**, *120*, 8834–8847.

(20) Kveseth, K.; Seip, R.; Kohl, D. A. *Acta. Chem. Scand. Ser. A* **1980**, *34*, 31–342.

(21) Caminati, W.; Grassi, G.; Bauder, A. *Chem. Phys. Lett.* **1988**, *148*, 13–16.

(22) Fincher, C. R.; Chen, C. E.; Heeger, A. J.; MacDiarmid, A. G.; Hastings, J. B. *Phys. Rev. Lett.* **1981**, *48*, 100–104.

(23) Yannoni, C. S.; Clarke, T. C. *Phys. Rev. Lett.* **1983**, *51*, 1191–1193.

(24) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

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

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

(27) Jacquemin, D.; Preat, J.; Wathelet, V.; Perpete, E. A. *THEOCHEM* **2005**, *731*, 67–72.

(28) Ciofini, I.; Adamo, C.; Chermette, H. *J. Chem. Phys.* **2005**, *123*, 121102.

(29) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 10478–10486.

(30) Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.

(31) Sancho-Garcia, J. C.; Perez-Jimenez, A. J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5874–5879.

(32) Peach, M. J. G.; Tellgren, E.; Salek, P.; Helgaker, T.; Tozer, D. J. *J. Phys. Chem. A* **2007**, *111*, 11930–11935.

(33) Chabbal, S.; Jacquemin, D.; Adamo, C.; Stoll, H.; Leininger, T. *J. Chem. Phys.* **2010**, *133*, 151104.

(34) Jacquemin, D.; André, J. M.; Perpète, E. A. *J. Chem. Phys.* **2004**, *121*, 4389–4396.

(35) Jacquemin, D.; Perpete, E. A.; Chermette, H.; Ciofini, I.; Adamo, C. *Chem. Phys.* **2007**, *332*, 79–85.

(36) Chabbal, S.; Stoll, H.; Werner, H. J.; Leininger, T. *Mol. Phys.* **2010**, *108*, 3373–3382.

(37) Jacquemin, D.; Femenias, A.; Chermette, H.; André, J. M.; Perpète, E. A. *J. Phys. Chem. A* **2005**, *109*, 5734–5741.

(38) Jacquemin, D.; Femenias, A.; Chermette, H.; Ciofini, I.; Adamo, C.; André, J. M.; Perpète, E. A. *J. Phys. Chem. A* **2006**, *110*, 5952–5959.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian Inc.: Wallingford, CT, 2009.

(40) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095–9102.

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

(42) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.

(43) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys.* **2007**, *126*, 124115.

(44) Neese, F. *ORCA*; Universität Bonn: Bonn, Germany, 2008.

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

(46) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(47) Yang, W.; Wu, Q. *Phys. Rev. Lett.* **2002**, *89*, 143002.

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

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

(50) Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

CT1006532

JCTC Journal of Chemical Theory and Computation

Average Local Ionization Energies as a Route to Intrinsic Atomic Electronegativities

Peter Politzer,*,† Zenaida Peralta-Inga Shields,† Felipe A. Bulat,‡ and Jane S. Murray†

CleveTheoComp LLC, 1951 West 26th Street, Suite 409, Cleveland, Ohio 44113, United States and Fable Theory & Computation LLC, P.O. Box 21811, Washington, D.C. 20009, United States

Received November 12, 2010

Abstract: Historically, two important approaches to the concept of electronegativity have been in terms of: (a) an atom in a molecule (e.g., Pauling) and (b) the chemical potential. An approximate form of the latter is now widely used for this purpose, although it includes a number of deviations from chemical experience. More recently, Allen introduced an atomic electronegativity scale based upon the spectroscopic average ionization energies of the valence electrons. This has gained considerable acceptance. However it does not take into account the interpenetration of valence and low-lying subshells, and it also involves some ambiguity in enumerating d valence electrons. In this paper, we analyze and characterize a formulation of relative atomic electronegativities that is conceptually the same as Allen's but avoids the aforementioned problems. It involves the property known as the average local ionization energy, $\bar{I}(\mathbf{r})$, defined as $\bar{I}(\mathbf{r}) = \sum\rho_i(\mathbf{r})|\varepsilon_i| / \rho(\mathbf{r})$, where $\rho_i(\mathbf{r})$ is the electronic density of the i^{th} orbital, having energy ε_i, and $\rho(\mathbf{r})$ is the total electronic density. $\bar{I}(\mathbf{r})$ is interpreted as the average energy required to remove an electron at the point $\mathbf{r}$. When $\bar{I}(\mathbf{r})$ is averaged over the outer surfaces of atoms, taken to be the 0.001 au contours of their electronic densities, a chemically meaningful scale of relative atomic electronegativities is obtained. Since the summation giving $\bar{I}(\mathbf{r})$ is over all occupied orbitals, the issues of subshell interpenetration and enumeration of valence electrons do not arise. The procedure is purely computational, and all of the atoms are treated in the same straightforward manner. The results of several different Hartree−Fock and density functional methods are compared and evaluated; those produced by the Perdew−Burke−Ernzerhof functional are chemically the most realistic.

Electronegativity

The concept of electronegativity is an old one, dating back at least to Berzelius in 1835,[1,2] and it is an extremely important one. It has been used to rationalize and predict a great deal of chemical behavior. However electronegativity is not a physical observable and cannot be determined experimentally. It is an arbitrarily defined property and therefore can be—and has been—formulated in many different ways. In this paper, after briefly examining some of these approaches, we shall focus upon ionizaton energy as a pathway to electronegativity.

Pauling assigned relative electronegativities to the elements on the basis of the estimated degrees of ionic character in the heteronuclear covalent bonds that they form.[3,4] His scale remains, in updated form,[5] a standard against which others are measured. Pauling viewed electronegativity as "the power of an atom in a molecule to attract electrons to itself".[6] His focus upon polarity and upon the atom in a molecular environment is reflected in many of the treatments of electronegativity that followed, different as they might be in detail. For example, Mulliken defined the electronegativity

* Corresponding author. E-mail: ppolitze@uno.edu.
† CleveTheoComp LLC.
‡ Fable Theory & Computation LLC.

10.1021/ct1006554
Published on Web 01/04/2011

χ in terms of the energetics of the atom in a molecule gaining or losing an electron:[7,8]

$$\chi = 0.5(I_{vs} + A_{vs}) \quad (1)$$

I_{vs} and A_{vs} are the valence-state ionization energy and electron affinity of the atom.

Allred and Rochow represented electronegativity as the electrostatic force exerted by the atom's effective nuclear charge upon an electron at its covalent radius.[9] Iczkowski and Margrave,[10] following earlier work by Pritchard and Sumner,[11] expressed electronegativity as $(\partial E/\partial Q)_{Q=0}$, where E and Q are the energy and the net charge of the atom *in a molecule*.

Already in 1961, the number and the diversity of electronegativity definitions was such that Iczkowski and Margrave were led to remark that "... there is some confusion as to what physical picture corresponds to the term electronegativity".[10] They pointed out that there was not even agreement as to its units, which had included energy, energy$^{1/2}$, force, force/distance, and potential. Numerous comparisons and critiques of the various formulations of electronegativity have appeared over the years.[2,12−19]

A new chapter in the electronegativity saga began in 1978. In the density functional treatment of a ground-state N-electron system having electronic density $\rho(\mathbf{r})$, the term "chemical potential" is applied to the Lagrangian multiplier μ used in minimizing the energy functional $E[\rho(\mathbf{r})]$ subject to the constraint of constant N, $N = \int \rho(\mathbf{r})d\mathbf{r}$.[16] The chemical potential can be expressed as

$$\mu = \left(\frac{\partial E[\rho]}{\partial \rho}\right)_{v(\mathbf{r})} \quad (2)$$

in which $v(\mathbf{r})$ is the external potential, which usually refers to the nuclei of the system. Parr et al. rewrote eq 2 as

$$\mu = \left(\frac{\partial E[N]}{\partial N}\right)_{v(\mathbf{r}),N_0} \quad (3)$$

where N_0 is the number of electrons in the ground state. The quantity μ was interpreted as the negative of the electronegativity χ:[20]

$$\mu = -\chi \quad (4)$$

The μ (and χ) defined by eqs 3 and 4 have been described as measuring the escaping tendency of an electron within the system.[16] This is significantly different from the conception of electronegativity held by Pauling (see above) and by many chemists.

Several points can be mentioned in support of eqs 3 and 4:

(a) It has been shown that μ must be uniform throughout a system at equilibrium.[20,21] This is consistent with Sanderson's postulate of electronegativity equalization.[22,23]

(b) Eq 2 can readily be converted (although not rigorously) to the easily evaluated form:

$$\mu = -0.5(I + A) = -\chi \quad (5)$$

where I and A are the system's ionization potential and electron affinity. This can be done by: (a) assuming that E is a quadratic function of N, (b) applying a finite-difference approximation to the derivative $(\partial E/\partial N)_{v(\mathbf{r}),N_0}$, or (c) expanding $E(N)$ in a Taylor series around N_0 and truncating after the second-order term.[24−26] The point is that eq 5 has the same form as the electronegativity expression introduced by Mulliken, eq 1.[7,8] However Mulliken's equation pertains to the valence state of the atom, and eq 5 pertains to the ground state.

(c) The electronic chemical potential as given by eq 3 is clearly analogous to the chemical potential of a component i of a macroscopic system in classical thermodynamics, which must also be uniform at equilibrium.

As was already mentioned, the approach to electronegativity represented by eqs 3 and 4 differs fundamentally from that of Pauling. The latter focuses upon the degree of ionic character of the atom in a molecule, while eqs 3 and 4 are for the ground state and can refer to a molecule as well as an atom. The differences are sufficiently important that both Pearson[27] and Allen[28,29] have suggested that Pauling's electronegativity and the chemical potential be regarded as two separate and distinct properties.

Eqs 3 and 4 have also been criticized on various grounds,[2,29,30] such as the validity of taking E to be a differentiable function of N given that N is restricted to having integral values. This issue has been discussed on a number of occasions.[16,31−33] Note that this problem does not arise with the thermodynamic chemical potential, which involves differentiating with respect to the number of moles of component i; this can certainly have nonintegral values.

If one accepts the differentiability of $E(N)$ and subsequently arrives at the approximation represented by eq 5, then it should be recognized that eq 3, and therefore eq 5, requires a constant nuclear potential. Thus the I and A in eq 5 should be the vertical values, not the adiabatic. For atoms this is of course trivial, but for molecules the effect can be significant, especially for A.[34] Furthermore, the vertical electron affinity is often negative for closed-shell molecules.[35] Then $E(N)$ must have a minimum for N in the vicinity of N_0; by eq 3, such molecules have zero or near-zero chemical potentials (as is predicted for any molecule by Thomas−Fermi theory).[16]

In practice, the requirement of constant $v(\mathbf{r})$ is frequently ignored, and I and A are taken to be the ground-state adiabatic values rather than the vertical. Thus, we are confronted with three similar formulas for χ:

$$\chi = 0.5(I_{vs} + A_{vs})_{N_0} \quad (1a)$$

$$\chi = 0.5(I_{vert} + A_{vert})_{N_0} \quad (6)$$

$$\chi = 0.5(I_{adiab} + A_{adiab})_{N_0} \quad (7)$$

Eq 1 is due to Mulliken and pertains to the atom in its valence state in the molecule. Eq 6 is based approximately upon eqs 3 and 4 and obeys the restriction that $v(\mathbf{r})$ must be constant. Eq 7 is also based approximately upon eqs 3 and

Table 1. Some Literature Electronegativity Values. Note that Pauling's are in Arbitrary Units

Atom Eq. (7); eV[a] Pauling; arbitrary units[b] Configuration energies; eV[c]							
Main group elements							
H 7.18 2.20 13.61							**He** --- --- 24.59
Li 3.01 0.98 5.392	**Be** 4.9 1.57 9.323	**B** 4.29 2.04 12.13	**C** 6.27 2.55 15.05	**N** 7.30 3.04 18.13	**O** 7.54 3.44 21.36	**F** 10.41 3.98 24.80	**Ne** --- --- 28.31
Na 2.85 0.93 5.140	**Mg** 3.75 1.31 7.646	**Al** 3.23 1.61 9.539	**Si** 4.77 1.90 11.33	**P** 5.62 2.19 13.33	**S** 6.22 2.58 15.31	**Cl** 8.30 3.16 16.97	**Ar** --- --- 19.17
K 2.42 0.82 4.340	**Ca** 2.2 1.00 6.113	**Ga** 3.2 1.81 10.39	**Ge** 4.6 2.01 11.80	**As** 5.3 2.18 13.08	**Se** 5.89 2.55 14.34	**Br** 7.59 2.96 15.88	**Kr** --- --- 17.54
First transition series							

Sc	**Ti**	**V**	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Cu**	**Zn**
3.34	3.45	3.6	3.72	3.72	4.06	4.3	4.40	4.48	4.45
1.36	1.54	1.63	1.66	1.55	1.83	1.88	1.91	1.90	1.65
7.042	8.170	9.063	9.77	10.34	10.64	10.86	11.13	10.96	9.395

[a] Taken from ref 16. [b] Taken from ref 36. Note that Pauling's are in arbitrary units. [c] Taken from refs 41 and 42.

4 but disregards the limitation upon $v(\mathbf{r})$. Eqs 6 and 7 are intended to apply to molecules as well as to atoms; for the latter, they are equivalent.

Eq 7, despite the problems mentioned above, is now widely used to calculate atomic and molecular electronegativities. How meaningful are they, from a chemical standpoint? In Table 1 are the results obtained with eq 7 for the elements H−Kr, using experimental ground-state I and A.[16] In general, they follow the expected trends in the periodic table, increasing from left to right in the horizontal rows and decreasing from top to bottom in the vertical columns. However there are a number of deviations from chemical experience (and from Pauling's electronegativities,[36] also in Table 1); some of the more striking ones are Cl > O, Cl > N, Br ∼ O, Br > N, H > C, H > S, H ∼ N. Thus, for example, amines would not be predicted to form hydrogen bonds!

These unrealistic relative values can usually be avoided by utilizing eq 1, with valence-state ionization energies and electron affinities.[10,18,37−39] However this produces a different χ for each valence state of an atom, of which there can be several; for instance, Hinze and Jaffé list seven possible valence states for triply coordinated nitrogen.[37] There can be serious ambiguity in specifying the valence state.[38,39] An analogous problem is actually inherent in any electronegativity treatment that focuses upon the atom in a molecule, including Pauling's, since a given atom differs somewhat from one molecular environment to another. (In this context, see Allen.)[29] It can be argued that there is a need for an electronegativity definition that is intrinsic to an atom and yet serves as a realistic guide to its interactions with others.

Electronegativity and Ionization Energy

The ionization energies of ground-state atoms are considerably larger than their electron affinities, often by factors of 5−10.[36] In eq 7, it is therefore I_{adiab} that primarily governs the magnitude of χ. Sacher and Currie used a double linear regression to determine the combination of I_{adiab} and A_{adiab} that best correlates with the Allred−Rochow electronegativities and concluded that the contribution of A_{adiab} is insignificant.[40] However using I_{adiab} alone to represent χ, i.e.,

$$\chi = I_{\mathrm{adiab}} \tag{8}$$

is also beset with problems O < N, H ∼ O, H > Cl, H > Br, H > C, etc.[36]

Allen et al. have introduced an electronegativity scale that is within the context of ionization energy but does not lead to the chemically unrealistic predictions that result from eqs 7 and 8.[1,15,29,41,42] They proposed that electronegativity be defined as the "configuration energy" (CE) of the ground-state free atom. By configuration energy, they mean the average ionization energy of its valence electrons. Thus, for the main group (nontransition) elements:

$$\chi = \mathrm{CE} = \frac{n_s\varepsilon_s + n_p\varepsilon_p}{n_s + n_p} \tag{9}$$

In eq 9, n_s and n_p are the numbers of s and p valence electrons, and ε_s and ε_p are the multiplet-averaged differences in total energy between the ground-state neutral atom and its monopositive ion resulting from the loss of a s or p electron. Spectroscopic data are used to obtain ε_s and ε_p insofar as possible; indeed Allen originally referred to the CE as "spectroscopic electronegativities."[15] The quantities ε_s and ε_p can also be approximated computationally, using the appropriate orbital energies.[15]

The configuration energy concept has been quite effective in producing chemically meaningful electronegativities (Table 1), as discussed in detail by Allen.[1,15,29] It has gained considerable acceptance over the past 20 years.[41]

The approach has also been extended to the transition series, but for these it is not as straightforward as for the main group elements.[15,42] For example, there is some arbitrariness in deciding the numbers of d electrons to be used in calculating the configuration energies.

For both the main group and the transition elements, there is furthermore the issue that interpenetration between the valence shell and lower-lying subshells is not being taken into account. This can be quite significant, even for main group elements and certainly for those in the transition series. A quantitative measure of this can be obtained by computing the exchange/repulsion interaction energy between a valence electron and one in an inner subshell and by comparing this to the expected value in the absence of interpenetration.[43] For the bromine 4p−3s, 3p combination, the ratio (actual/no penetration) is about 0.91; for the sulfur 3p−2s, 2p, it is 0.92. The effect is much greater for 3d electrons; for 3d−3s, 3p interactions, the ratio is only 0.7 to 0.8!

In an earlier preliminary study,[44] we proposed a variation of Allen's approach that is conceptually the same but that avoids the problems associated with subshell interpenetration and enumeration of d valence electrons. We now present a

detailed characterization of this procedure. It involves the average local ionization energies of the atoms.

Average Local Ionization Energy

The property known as the average local ionization energy, $\bar{I}(\mathbf{r})$, was introduced in 1990 as a measure of the energy required to remove an electron from a specific point $\mathbf{r}$ in an atom or molecule.[45] The focus is upon the point in space, not a particular orbital. $\bar{I}(\mathbf{r})$ is given by

$$\bar{I}(\mathbf{r}) = \frac{\sum_i \rho_i(\mathbf{r})|\varepsilon_i|}{\rho(\mathbf{r})} \quad (10)$$

in which $\rho_i(\mathbf{r})$ is the electronic density of orbital $\varphi_i(\mathbf{r})$, having energy ε_i, and $\rho(\mathbf{r})$ is the total electronic density. The summation is over all occupied orbitals.

The interpretation of $\bar{I}(\mathbf{r})$ as a local ionization energy is formally justifiable in Hartree−Fock theory if it is assumed that the loss of an electron from one orbital does not affect the others; some support for this is provided by Koopmans' theorem.[46,47] $\bar{I}(\mathbf{r})$ as defined by eq 10 has also been shown to be effective within the framework of Kohn−Sham density functional methodology.[48,49] The magnitudes are different from the Hartree−Fock but the relative values and trends are generally the same, which is the key point. This will be addressed again in the next section.

The lowest values of $\bar{I}(\mathbf{r})$ indicate the locations of the least tightly held, most reactive electrons. $\bar{I}(\mathbf{r})$ has indeed been found to be quite successful in predicting and ranking sites for electrophilic attack as well as pK_a values.[45,48−52] In these studies, $\bar{I}(\mathbf{r})$ is typically computed on the molecular "surface", which is usually taken to be the 0.001 au (electrons/bohr^3) contour of the electronic density $\rho(\mathbf{r})$, as proposed by Bader et al.[53]

The significance of $\bar{I}(\mathbf{r})$ is not limited to reactive behavior. It has been shown to be linked to local kinetic energy density, atomic shell structure, and local polarizability/hardness. It can be used to identify radical sites and strained C−C bonds. The various aspects of $\bar{I}(\mathbf{r})$ are discussed in detail elsewhere.[54,55]

Average Local Ionization Energy and Electronegativity

It has been demonstrated that the 0.001 au surfaces of the atoms lie within their valence shells.[56,57] Thus if we compute the average value of $\bar{I}(\mathbf{r})$ on these surfaces, we are in fact obtaining primarily the average ionization energies of their valence electrons−the same concept as Allen's configuration energies![41,42] However if an inner orbital $\varphi_i(\mathbf{r})$ has a significant presence in the valence shell, then this will be reflected in its $\rho_i(\mathbf{r})$ on the 0.001 au surface, and its contribution to $\bar{I}(\mathbf{r})$ will be included. Thus the problem of accounting for subshell interpenetration does not arise, because the summation in eq 10 is over all of the atom's electrons. Similarly, the need to enumerate d valence electrons is eliminated, again because of summing over all electrons. These two issues that are associated with Allen's approach are accordingly resolved.

We shall use $\bar{I}_{S,ave}$ to designate the average value of $\bar{I}(\mathbf{r})$ on the 0.001 au surface of an atom. In our earlier study,[44] we computed $\bar{I}_{S,ave}$ for the atoms H−Kr. Clementi's Hartree−Fock wave functions, written in terms of extended basis sets of Slater-type orbitals, were used for He−Kr;[58] hydrogen was treated exactly. The resulting $\bar{I}_{S,ave}$ correlated well with Allen's configuration energies and showed the relative values and trends expected for electronegativity, with two exceptions:[44] $\bar{I}_{S,ave}$ for hydrogen was larger than for sulfur, 13.61 vs 13.26 eV, and sulfur in turn was less than carbon, 14.30 eV. These are contrary to chemical experience, e.g., the known formation of −S−H···X hydrogen bonds.[59]

Clementi's wave functions were at the Hartree−Fock level and therefore did not include electronic correlation. Furthermore, the p and d subshells were treated as spherically symmetrical. In order to assess the effects of these factors, and to more fully characterize the $\bar{I}_{S,ave}$ approach to electronegativity, we now present the $\bar{I}_{S,ave}$ for the atoms H−Kr as computed by several different approaches:

(a) Hartree−Fock, STO: Extended basis sets of Slater-type orbitals, p and d subshells spherically symmetrical.[58] The resulting $\bar{I}_{S,ave}$ are the ones given in our earlier paper.[44]
(b) Optimized Potential Method (OPM): Exchange-only, spherically symmetrical Kohn−Sham. Formally the same energy functional as Hartree−Fock, OPM is used to obtain Kohn−Sham exchange potentials.[60,61]
(c) Hartree−Fock, 6-311G(3d, 2p).
(d) PBEPBE/6-311G(3d, 2p): Pure density functional,[62] no Hartree−Fock exchange.
(e) B3LYP/6-311G(3d, 2p): Three-parameter hybrid functional,[63,64] 20% Hartree−Fock exchange.
(f) MO62X/6-311G(3d, 2p): Hybrid meta density functional,[65,66] 54% Hartree−Fock exchange.

Procedures (d−f) are Kohn−Sham, with local exchange and correlation plus various amounts of Hartree−Fock exchange. In (c−f), the p and d subshells are not necessarily spherically symmetrical; the orbitals have the occupancies corresponding to the ground-state configurations. The wave functions for (c−f) were obtained with Gaussian 09,[67] while the $\bar{I}_{S,ave}$ were computed using the WFA surface analysis suite.[68] For the PBEPBE functional, we also investigated the addition of diffuse functions to the basis set but found these to have no significant effect.

The results, for procedures (a−f), are in Table 2. To facilitate comparisons, Table 3 lists their relative values, based upon fluorine being assigned a value of 4.00 in each case. (This was its original Pauling electronegativity.)[3,4]

Discussion

Tables 2 and 3 show that the overall variation of $\bar{I}_{S,ave}$ is as expected of electronegativity: it increases from left to right in the horizontal rows and decreases from top to bottom in the vertical columns. This is depicted graphically in Figure 1.

Looking first at the absolute values of $\bar{I}_{S,ave}$, in Table 2, it is evident that they fall into two groups: The Hartree−Fock and the OPM (exchange-only Kohn−Sham) $\bar{I}_{S,ave}$ have the larger magnitudes, the exchange plus correlation Kohn−Sham

Table 2. Unscaled $\bar{I}_{S,ave}$ Values, in eV, for Main Group Elements and First Transition Series

Atom
HF-STO, from ref. 44
OPM
HF/6-311G(3d,2p)
PBEPBE/6-311G(3d,2p)
B3LYP/6-311G(3d,2p)
M06-2X/6-311G(3d,2p)

H							**He**
13.61							24.98
13.60							24.94
13.60							24.95
7.55							15.63
8.75							17.91
10.31							20.57
Li	**Be**	**B**	**C**	**N**	**O**	**F**	**Ne**
5.34	8.42	11.12	14.30	17.87	19.48	21.89	24.97
5.34	8.40	11.33	13.45	15.94	18.57	21.45	24.56
5.34	8.41	11.47	14.38	17.44	19.27	21.84	25.25
3.22	5.60	6.99	8.39	9.98	10.91	12.37	14.42
3.65	6.31	8.07	9.79	11.63	12.84	14.59	16.94
4.20	7.26	9.31	11.50	13.89	15.29	17.42	20.27
Na	**Mg**	**Al**	**Si**	**P**	**S**	**Cl**	**Ar**
5.01	6.89	7.98	9.75	12.12	13.26	15.05	17.30
5.01	6.87	7.88	9.08	10.71	12.59	14.66	16.89
5.03	6.88	8.47	10.11	12.21	13.54	15.08	17.23
3.15	4.70	5.41	6.21	7.52	8.45	9.52	11.00
3.60	5.29	6.08	7.02	8.48	9.57	10.80	12.46
4.28	6.18	7.06	8.23	9.99	11.18	12.58	14.47
K	**Ca**	**Ga**	**Ge**	**As**	**Se**	**Br**	**Kr**
5.20	5.34	8.12	9.47	11.40	12.18	13.53	15.31
5.12	5.34	8.01	8.77	10.03	11.56	13.11	14.84
5.12	5.34	8.94	9.88	11.40	12.32	13.67	15.34
3.31	3.78	5.84	6.07	6.93	7.67	8.69	9.94
3.71	4.22	6.60	6.89	7.83	8.70	9.84	11.21
4.22	4.98	7.41	8.10	9.29	10.25	11.51	13.00

Sc	**Ti**	**V**	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Cu**	**Zn**
5.82	6.19	6.49	6.79	6.96	7.26	7.51	7.77	7.98	8.17
5.78	6.08	6.42	6.73	7.00	7.26	7.51	7.73	7.01	8.17
5.70	5.96	6.20	5.65	5.67	5.76	5.84	5.93	7.58	7.77
3.99	4.22	4.38	3.56	4.66	4.76	3.77	3.83	5.33	5.45
4.52	4.75	4.95	4.07	5.29	4.27	4.36	5.90	5.43	6.22
5.25	5.48	5.67	4.88	6.00	6.13	6.26	6.40	6.57	6.72

have the smaller. This is fully consistent with earlier work.[55] The two sets of Hartree−Fock results are in general quite similar for the main group elements but differ by more than 1.0 eV for the transition atoms Cr−Ni. This suggests that symmetry, or lack thereof, is more important in the d subshell than in the p.

The fact that the OPM $\bar{I}_{S,ave}$ is overall quite close to the Hartree−Fock provides significant support for the validity of extending the definition of $\bar{I}(\mathbf{r})$, eq 10, to the Kohn−Sham framework. The OPM energy functional is formally the same as the Hartree−Fock; the two methods differ only in the exchange potential that enters the single-particle Hartree−Fock or Kohn−Sham equations. In the latter, this potential is constrained to be local (and multiplicative), whereas in the former it is nonlocal. In Hartree−Fock theory, the interpretation of $\bar{I}(\mathbf{r})$ as an average local ionization energy is partially justified by Koopmans' theorem,[46,47] but the Kohn−Sham counterpart, Janak's theorem,[69] holds only for the highest-occupied and lowest-unoccupied orbitals. The OPM and Hartree−Fock results in Table 2 show that when the energy functionals are formally identical, the $\bar{I}_{S,ave}$ obtained with the local and nonlocal potentials not only show the same trends but also have quite similar values.

It was mentioned earlier in this paper that the widely used electronegativity expression, eq 7, produces a number of chemically unrealistic results, including Cl > O, Cl > N, Br ∼ O, Br > N, H > C, H > S, and H ∼ N. (See Table 1.) Do any of these persist in the $\bar{I}_{S,ave}$ in Table 2? The Hartree−Fock procedures as well as the OPM do show H > S; the OPM also yields H > C (Table 2). Furthermore, the Hartree−Fock and OPM methods encounter a problem with C, S, and Br, predicting that C > S and C > Br. The exchange plus correlation approaches fare better, but only the PBEPBE functional is really satisfactory with respect to the relative C, S, and Br values. In particular, it is the only one of the six methods tested that shows bromine to be distinctly more electronegative than carbon.

Is it the complete absence of Hartree−Fock exchange that makes the PBEPBE functional more effective than the B3LYP and the M06-2X for present purposes? To test this possibility, we computed $\bar{I}_{S,ave}$ for a series of atoms (H, C, N, O, S, Cl, Br) using the M06-L functional,[65,66] which also involves no Hartree−Fock exchange, and the 6-311G(3d,2p) basis set. The results for this group of atoms are overall realistic, and $\bar{I}_{S,ave}$ is indeed larger for bromine than for carbon, although by relatively little, 8.99 vs 8.82 eV.

Table 3. Scaled $\bar{I}_{S,ave}$ Values for Main Group Elements and First Transition Series[a]

Atom HF-STO OPM HF/6-311G(3d,2p) PBEPBE/6-311G(3d,2p) B3LYP/6-311G(3d,2p) M06-2X/6-311G(3d,2p)							
H 2.49 2.54 2.49 2.44 2.40 2.37							**He** 4.57 4.65 4.57 5.06 4.91 4.72
Li 0.98 1.00 0.98 1.04 1.00 0.97	**Be** 1.54 1.57 1.54 1.81 1.73 1.67	**B** 2.03 2.11 2.10 2.26 2.21 2.14	**C** 2.61 2.51 2.64 2.71 2.68 2.64	**N** 3.27 2.97 3.20 3.23 3.19 3.19	**O** 3.56 3.46 3.53 3.53 3.52 3.51	**F** 4.00 4.00 4.00 4.00 4.00 4.00	**Ne** 4.56 4.58 4.63 4.66 4.65 4.65
Na 0.92 0.94 0.92 1.02 0.99 0.98	**Mg** 1.26 1.28 1.26 1.52 1.45 1.42	**Al** 1.46 1.47 1.55 1.75 1.67 1.62	**Si** 1.78 1.69 1.85 2.01 1.93 1.89	**P** 2.22 2.00 2.24 2.43 2.32 2.29	**S** 2.42 2.35 2.48 2.73 2.63 2.57	**Cl** 2.75 2.73 2.76 3.08 2.96 2.89	**Ar** 3.16 3.15 3.16 3.56 3.42 3.32
K 0.95 0.96 0.94 1.07 1.02 0.97	**Ca** 0.98 1.00 0.98 1.22 1.16 1.14	**Ga** 1.48 1.50 1.64 1.89 1.81 1.70	**Ge** 1.73 1.64 1.81 1.96 1.89 1.86	**As** 2.08 1.87 2.09 2.24 2.15 2.13	**Se** 2.23 2.16 2.26 2.48 2.39 2.35	**Br** 2.47 2.45 2.51 2.81 2.70 2.64	**Kr** 2.80 2.77 2.81 3.22 3.07 2.98

Sc	**Ti**	**V**	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Cu**	**Zn**
1.06	1.13	1.19	1.24	1.27	1.33	1.37	1.42	1.46	1.49
1.08	1.13	1.20	1.26	1.31	1.35	1.40	1.44	1.31	1.52
1.05	1.09	1.14	1.04	1.04	1.06	1.07	1.09	1.39	1.42
1.29	1.37	1.42	1.15	1.51	1.54	1.22	1.24	1.72	1.76
1.24	1.30	1.36	1.12	1.45	1.17	1.20	1.62	1.49	1.71
1.21	1.26	1.30	1.12	1.38	1.41	1.44	1.47	1.51	1.54

[a] Fluorine is assigned 4.00.

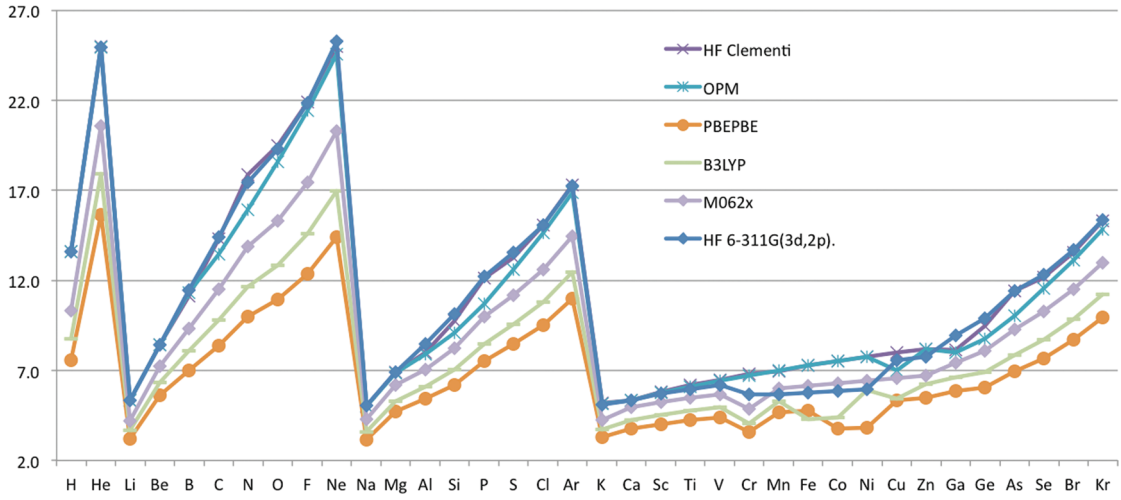


Figure 1. Computed unscaled $\bar{I}_{S,ave}$ values, in eV, for main group elements and first transition series.

Scaling the $\bar{I}_{S,ave}$ so that the value for fluorine is exactly 4.00 by each method considerably diminishes the differences between the six sets of results (Table 3). The scaled PBEPBE, B3LYP, and M06-2X $\bar{I}_{S,ave}$ of the main group atoms usually differ very little. The method dependence is greater for the first transition series, but it is not really clear what should be expected. Perhaps the key points for the transition elements are that the magnitudes tend to be rather small and to vary relatively little over the whole series (about 0.4), for each of the three procedures.

The PBEPBE functional gives the lowest results of all of the methods, followed by the B3LYP, and the M06-2X. This

reflects the amount of Hartree−Fock nonlocal exchange in each of these functionals: 0% for PBEPBE, 20% for B3LYP, and 54% for M06-2X.

Summary

The average local ionization energies on the 0.001 au surfaces of the ground-state atoms, as computed with the PBEPBE exchange plus correlation density functional, provide an effective measure of their relative electronegativities. This is an alternative means of implementing the concept introduced by Allen et al.[1,15,29,41,42] This approach is purely computational and treats all atoms in exactly the same straightforward manner. The problem of accounting for interpenetration between valence and lower-lying subshells does not arise nor does the need to enumerate d valence electrons.

It should be noted that we are addressing electronegativity, not the chemical potential. We do not necessarily assume the validity of eq 4, $\mu = -\chi$. As mentioned earlier, Pearson[27] and Allen[28,29] questioned this already some years ago. We agree with Allen that electronegativity should be an intrinsic property of a ground-state atom[29] as opposed to an atom in a molecule or in some valence state, given the ambiguity that these entail.

If we do not require that $\mu = -\chi$, then the evidence[20,21] that μ must be uniform throughout a system at equilibrium is not relevant for χ. While the notion of electronegativity equalization may be appealing, it also seems counter to chemical experience, as pointed out by Allen[29] and by Hinze.[2] Atoms do retain much of their identities in molecules; this can be confirmed by looking at plots of molecular electronic densities.[70,71] The rearrangements of electrons that accompany the formation of a molecule are very subtle, as can be seen in density difference plots.[70−72] The free atom electronegativities provide us with some initial qualitative guidelines concerning these rearrangements.

Acknowledgment. We thank Professor Donald G. Truhlar for very helpful discussions.

References

(1) Allen, L. C. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 2, pp 835−852.

(2) Hinze, J. In *Pauling's Legacy: Modern Modelling of the Chemical Bond*; Z. B. Maksic, Z. B., Orville-Thomas, W. J., Eds.; Elsevier: Amsterdam, The Netherlands, 1999; Chapter *7*, 189−212.

(3) Pauling, L. *J. Am. Chem. Soc.* **1932**, *54*, 3570–3582.

(4) Pauling, L.; Yost, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **1932**, *18*, 414–416.

(5) Allred, A. L. *J. Inorg. Nucl. Chem.* **1961**, *17*, 215–221.

(6) Pauling, L. *The Nature of the Chemical Bond*, 2nd ed.; Cornell University Press: Ithaca, NY, 1948, *58*.

(7) Mulliken, R. S. *J. Chem. Phys.* **1934**, *2*, 782–793.

(8) Mulliken, R. S. *J. Chem. Phys.* **1935**, *3*, 573–585.

(9) Allred, A. L.; Rochow, E. G. *J. Inorg. Nucl. Chem.* **1958**, *5*, 264–268.

(10) Iczkowski, R. P.; Margrave, J. L. *J. Am. Chem. Soc.* **1961**, *83*, 3547–3351.

(11) Pritchard, H. O.; Sumner, F. H. *Proc. Roy. Soc. (London)* **1956**, *A235*, 136–143.

(12) Pritchard, H. O.; Skinner, H. A. *Chem. Rev.* **1955**, *55*, 745–786.

(13) Ferreira, R. *Adv. Chem. Phys.* **1967**, *13*, 55–84.

(14) Mullay, J. *Struct. Bonding (Berlin)* **1987**, *66*, 1–25.

(15) Allen, L. C. *J. Am. Chem. Soc.* **1989**, *111*, 9003–9014.

(16) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(17) Sproul, G. D. *J. Phys. Chem.* **1994**, *98*, 6699–6703.

(18) Bergmann, D.; Hinze, J. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 150–163.

(19) Politzer, P.; Grice, M. E.; Murray, J. S. *J. Mol. Struct. (Theochem)* **2001**, *549*, 69–76.

(20) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. *J. Chem. Phys.* **1978**, *68*, 3801–3807.

(21) Politzer, P.; Weinstein, H. *J. Chem. Phys.* **1979**, *71*, 4218–4220.

(22) Sanderson, R. T. *Science* **1951**, *114*, 670–672.

(23) Sanderson, R. T. *J. Am. Chem. Soc.* **1952**, *74*, 272–274.

(24) Gazquez, J. L.; Ortiz, E. *J. Chem. Phys.* **1984**, *81*, 2741–2748.

(25) Politzer, P.; Huheey, J. E.; Murray, J. S.; Grodzicki, M. *J. Mol. Struct. (Theochem)* **1992**, *259*, 99–120.

(26) Politzer, P.; Murray, J. S. *Chem. Phys. Lett.* **2006**, *431*, 195–198.

(27) Pearson, R. G. *Acc. Chem. Res.* **1990**, *23*, 1–2.

(28) Allen, L. C. *Acc. Chem. Res.* **1990**, *23*, 175–176.

(29) Allen, L. C. *Int. J. Quantum Chem.* **1994**, *49*, 253–277.

(30) Nguyen-Dang, T. T.; Bader, R. F. W.; Essén, H. *Int. J. Quantum Chem.* **1982**, *22*, 1049–1058.

(31) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L., Jr. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.

(32) Zhang, Y.; Yang, W. *Theor. Chem. Acc.* **2000**, *103*, 346–348.

(33) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1874.

(34) Politzer, P.; Murray, J. S.; Concha, M. C.; Jin, P. *Collect. Czech. Chem. Commun.* **2007**, *72*, 51–63.

(35) Pearson, R. G. *Inorg. Chem.* **1988**, *27*, 734–740.

(36) Allred, A. L. *Inorg. Nucl. Chem.* **1961**, *17*, 215–221.

(37) Hinze, J.; Jaffé, H. H. *J. Am. Chem. Soc.* **1962**, *84*, 540–546.

(38) Liebman, J. F.; Huheey, J. E. *Phys. Rev. D* **1987**, *36*, 1559–1561.

(39) Bratsch, S. G. *J. Chem. Educ.* **1988**, *65*, 34–41, 223−227.

(40) Sacher, E.; Currie, J. F. *J. Electron Spectrosc. Relat. Phenom.* **1988**, *46*, 173–177.

(41) Mann, J. B.; Meek, T. L.; Allen, L. C. *J. Am. Chem. Soc.* **2000**, *122*, 2780–2783.

(42) Mann, J. B.; Meek, T. L.; Knight, E. T.; Capitani, J. F.; Allen, L. C. *J. Am. Chem. Soc.* **2000**, *122*, 5132–5137.

(43) Politzer, P.; Daiker, K. C. *Chem. Phys. Lett.* **1973**, *20*, 309–316.

(44) Politzer, P.; Murray, J. S.; Grice, M. E. *Collect. Czech. Chem. Commun.* **2005**, *70*, 550–558.

(45) Sjoberg, P.; Murray, J. S.; Brinck, T.; Politzer, P. *Can. J. Chem.* **1990**, *68*, 1440–1443.

(46) Koopmans, T. A. *Physica* **1934**, *1*, 104–113.

(47) Nesbet, R. K. *Adv. Chem. Phys.* **1965**, *9*, 321–363.

(48) Politzer, P.; Abu-Awwad, F.; Murray, J. S. *Int. J. Quantum Chem.* **1998**, *69*, 607–613.

(49) Politzer, P.; Murray, J. S.; Concha, M. C. *Int. J. Quantum Chem.* **2002**, *88*, 19–27.

(50) Murray, J. S.; Brinck, T.; Politzer, P. *J. Mol. Struct. (Theochem)* **1992**, *255*, 271–281.

(51) Brinck, T.; Murray, J. S.; Politzer, P. *Int. J. Quantum Chem.* **1993**, *48*, 73–88.

(52) Ma, Y.; Gross, K. C.; Hollingsworth, C. A.; Seybold, P. G.; Murray, J. S. *J. Mol. Model* **2004**, *10*, 235–239.

(53) Bader, R. W. F.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. *J. Am. Chem. Soc.* **1987**, *109*, 7968–7979.

(54) Politzer, P.; Murray, J. S. In *Theoretical Aspects of Chemical Reactivity*; Toro-Labbé, A., Ed.; Elsevier: Amsterdam, The Netherlands, 2007; Chapter *8*, 119−137.

(55) Politzer, P.; Murray, J. S.; Bulat, F. A. *J. Mol. Model.* **2010**, *16*, 173–1742.

(56) Politzer, P.; Murray, J. S.; Grice, M. E.; Brinck, T.; Ranganathan, S. *J. Chem. Phys.* **1991**, *95*, 6699–6704.

(57) Murray, J. S.; Politzer, P. *Croat. Chim. Acta* **2009**, *82*, 267–275.

(58) Clementi, E. *Tables of Atomic Functions*; IBM: San Jose, CA, 1965.

(59) Cotton, F. A.; Wilkinson, G. *Advanced Inorganic Chemistry*, 4th ed.; Wiley-Interscience: New York, 1980, *219*.

(60) Talman, J. D.; Shadwick, W. F. *Phys. Rev. A: At., Mol., Opt. Phys.* **1976**, *14*, 36–40.

(61) Heaton-Burgess, T.; Bulat, F. A.; Yang, W. *Phys. Rev. Lett.* **2007**, *98* (1−4), 256401.

(62) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(63) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.

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

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

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

(67) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(68) Bulat, F. A.; Toro-Labbé, A.; Brinck, T.; Murray, J. S.; Politzer, P. *J. Mol. Model.* **2010**, *16*, 1679–1691.

(69) Janak, J. F. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1978**, *18*, 7165–7168.

(70) Hazelrigg, M. J., Jr.; Politzer, P. *J. Phys. Chem.* **1969**, *73*, 1008–1011.

(71) Iwasaki, F.; Saito, Y. *Acta Crystallogr.* **1970**, *B26*, 251–260.

(72) Eisenstein, M.; Hirshfeld, F. L. *Chem. Phys.* **1979**, *42*, 465–474.

CT1006554

JCTC Journal of Chemical Theory and Computation

Highly Fluxional [Y(C(SiH(CH$_3$)$_2$)$_3$)$_3$]: A DFT Characterization of Structure and NMR Spectra

Matthias Lein*,†,‡ and John A. Harrison§

School of Chemical and Physical Sciences, Victoria University of Wellington, Kelburn Campus, Wellington 6140, New Zealand, Center for Theoretical Chemistry and Physics, New Zealand Institute for Advanced Study, Massey University Auckland, Private Bag 102904, North Shore City, Auckland 0745, New Zealand, and Institute of Natural Sciences, Massey University Auckland, Private Bag 102904, North Shore City, Auckland 0745, New Zealand

Received November 16, 2010

Abstract: The structure and NMR spectroscopic properties of [Y(C(SiH(CH$_3$)$_2$)$_3$)$_3$] are investigated with density functional theory calculations. The existence of a C_3 principal axis that was found experimentally is reproduced, but the calculations also find that the symmetry of the equilibrium structure of [Y(C(SiH(CH$_3$)$_2$)$_3$)$_3$] has to be reduced from the experimentally suggested C_{3v} or C_{3h} to C_3 in order to explain the observed Si*H* NMR chemical shifts. We show that the apparent mirror plane relating two agostic SiH(CH$_3$)$_2$ groups on each ligand is caused by the rapid interchange of the position of the third ligand, which could only be observed at much lower temperatures than used previously in the experiments.

Introduction

Since their discovery in the mid 1960s,[1] various types of agostic interactions have been described for many different systems. However, in some cases, an experimental verification of the presence of an agostic interaction is difficult because either reliable probes such as the characteristically lowered NMR coupling constants are unavailable for a particular compound or available methods such as structural information about CH-bond proximity to a metal center are not very accurate for such a prediction.[2] In these cases, computational inspection of the compound in question can often assist in the determination of the nature of a particular interaction.[3] The methods currently in use for such a computational analysis range from the theoretical reproduction of spectroscopic properties to the discussion of the bonding situation in terms of natural bonding orbitals or the topology of the electron density.[4]

In this study, we examine an yttrium complex with sterically bulky alkyl ligands. The compound was first synthesized by Sadow et al.,[5] who, on the basis of NMR data, described it as having six SiH agostic interactions. The high number of agostic contacts claimed in this molecule warrants a second look and a closer inspection of the situation of the three ligands in relation to the central metal atom.

The SiH agostic interactions reported are not unusual; in fact, several rare-earth complexes with silyl−amido agostic interactions had been reported before.[6] In those, the number of coordinating groups seems to depend on both electronic as well as steric effects. The larger [Eu(N(SiHMe$_2$)tBu)$_3$] complex coordinates through three β-agostic interactions,[6b] while the Me$_2$Si(C$_5$Me$_4$)$_2$YN(SiHMe$_2$)$_2$ ligand coordinates through a bis-β-agostic interaction[6c,e] and the bulkier Cp*$_2$YN(SiHMe$_2$)$_2$ ligand coordinates through a single β-agostic interaction.[6d]

In the initial characterization by Sadow et al., NMR spectroscopy was used as the main analytical tool in the investigation of the agostic nature of the SiH interaction. The authors could show that, in order to obtain sufficient resolution of the spectrum, the sample had to be cooled to approximately 190 K. Above this temperature, the relevant

* To whom correspondence should be addressed. E-mail: matthias.lein@vuw.ac.nz.

† Victoria University of Wellington.

‡ New Zealand Institute for Advanced Study, Massey University Auckland.

§ Institute of Natural Sciences, Massey University Auckland.

10.1021/ct100663y
Published on Web 12/30/2010

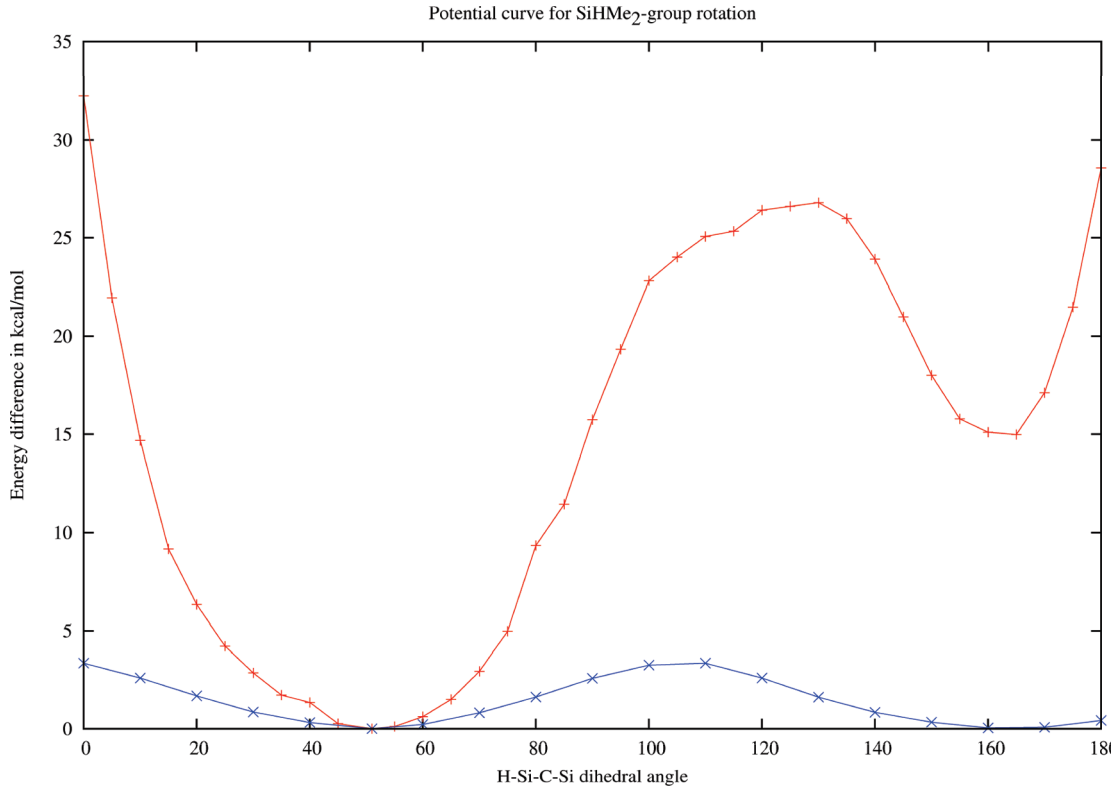


Figure 1. Energetic profile for the internal rotation of the noncoordinating $SiHMe_2$ group in **3**.

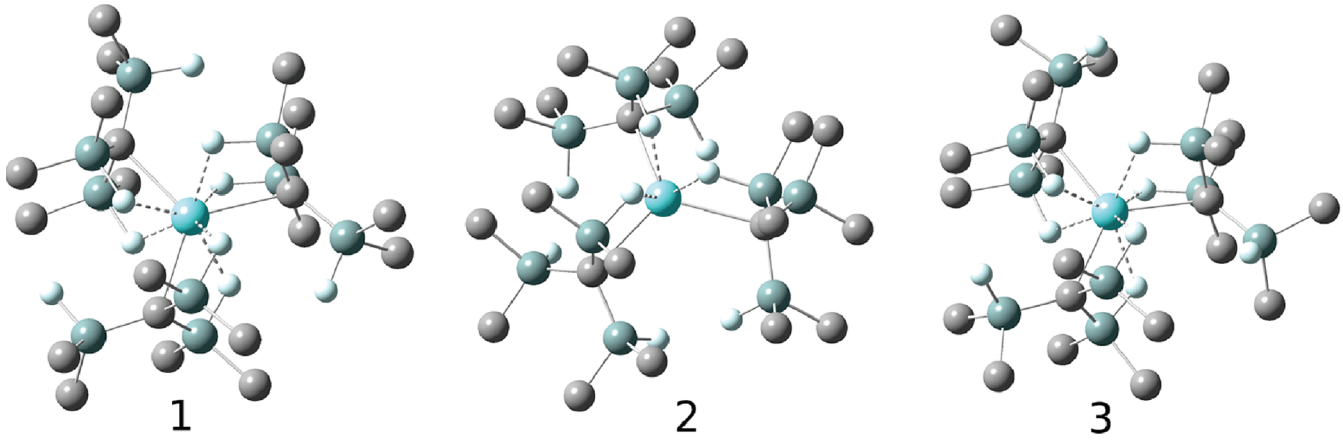


Figure 2. Structures of **1**, **2**, and **3** (only SiH hydrogen atoms are shown; all other hydrogen atoms are omitted for clarity). Bonds are shown as solid lines; Si−H···Y agostic interactions are shown as dashed lines.

SiH signals in the ^{1}H NMR spectrum coalesced into a single peak at 3.85 ppm.

Results and Discussion

Structures. In order to get a first insight into the structural properties of the yttrium compound in question, geometry optimizations were performed for both structures that had been previously suggested.[5] Those suggestions were based on the assumption of a pyramidalized structure (**1**), where the three $-C(Si(CH_3)_2H)_3$ ligands all coordinate from one side of the central yttrium atom, and a planar structure (**2**), where the coordinating carbon atoms of all three ligands and the central yttrium atom are in the same plane (see Figure 2). It was previously suggested[5] that a pyramidal structure might be the more likely one because $[Y(CH(SiMe_3)_2)_3]$, a similar compound, had been shown to be pyramidal.[7]

Both optimizations lead to stationary points on the potential energy hyper surface, but both structures, **2** with C_{3v} symmetry and structure **1** with C_{3h} symmetry, were calculated to be transition states. Unfortunately, it was also found that the C_{3v} structure is almost 70 kcal/mol higher in energy than the transition structure with C_{3h} symmetry and, in addition, has only three Si−H units coordinated to the central yttrium atom because of steric constraints, as opposed to six Si−H units, which had been suggested when this compound had been characterized for the first time (see Table 1). Consequently, the C_{3h} symmetric structure, **1**, was then used as a starting point for the search for the lowest energy

Table 1. Symmetries, Relative Energies (in kcal/mol), Number of Imaginary Modes (*i*), and ^{1}H NMR Chemical Shifts (in ppm) for Coordinating and Noncoordinating Si−H Protons

	symm	ΔE	*i*	δ_{coord}	$\delta_{noncoord}$
1	C_{3h}	10.6	3	6 × 3.45	3 × 5.26
1a	C_{3h}	14.0	3	6 × 3.43	3 × 5.13
2	C_{3v}	80.2	10	3 × 4.48	6 × 5.97
2a	C_s	45.9	4	2 × 4.01, 2 × 4.56	2 × 4.97, 2 × 4.84, 5.00
2b	C_s	13.3	2	2 × 5.12, 2 × 4.72, 4.23	2 × 5.03, 2 × 5.16
2c	C_3	8.9	0	3 × 3.70	3 × 4.97, 3 × 5.10
3	C_3	0.0	0	3 × 3.28, 3 × 3.70	3 × 4.75
3a	C_1	0.0	0	3.38, 3.45, 3.47, 3.52, 3.57, 3.62	4.76, 4.79, 4.80
3b	C_1	3.4	1		
exp[5]				6 × 3.40	3 × 4.71

conformation. In order to find such a structure, the coordinates of the transition state with C_{3h} symmetry were displaced according to the vibrational modes corresponding to imaginary frequencies, and the resulting initial structure with C_3 symmetry was again optimized. This resulted in a true minimum structure (**3**) with C_3 symmetry, which is 10.6 kcal/mol lower in energy than the previously obtained transition state (Figure 2). The retention of the C_3 axis is consistent with the spectral data obtained by Sadow and co-workers. In the energetic minimum, the three Si−H units that used to be coplanar with the horizontal mirror plane in the transition structure are now rotated out of the Y−C−Si plane by 52.7°. This leads to the observed lowering of symmetry by removal of the horizontal mirror plane. The other two Si(CH$_3$)$_2$H units of each of the three ligands are mostly unaffected by this internal rotation, and their mode of coordination to the central yttrium atom through their SiH groups does not change considerably, although a significant lengthening of the Y−H distances is noted. Furthermore, the loss of the horizontal mirror plane in **3** also makes the Si(CH$_3$)$_2$H units that used to be above the mirror plane different from the Si(CH$_3$)$_2$H units that used to be below the mirror plane. Those had been identical in the C_{3h} symmetrical transition structure, **1**, because of the σ_h-mirror symmetry between them. While the noncoordinating Si−H units showed a Y−H distance of 2.36 Å in the transition structure **1** and a much longer Y−H distance of 4.36 Å in the minimum structure **3** with C_3 symmetry, the corresponding changes in the groups with the coordinating Si−H units were much smaller. In those groups, the Y−H distances increased from 2.27 Å in the transition structure **1** to 2.30 Å and 2.36 Å in the minimum energy structure **3** for the groups that used to be above and below the mirror plane, respectively.

Considering the energetic gain of about 11 kcal/mol by the internal rotation of one Si(CH$_3$)$_2$H unit at each of the three C(Si(CH$_3$)$_2$H)$_3$ ligands at the same time by going from structure **1** to structure **3**, the question about the size of the energy barrier for the rotation of a single Si(CH$_3$)$_2$H unit around the C−Si(CH$_3$)$_2$H bond has to be asked. In order to determine this, two sets of surface scans were obtained and evaluated (see Figure 1). In the first scan, the Y−C−Si−H angle was varied from 0° to 180° in 5° increments with all other structural parameters unchanged. This constrained scan showed a large barrier for the rotation of the Si(CH$_3$)$_2$H unit of 26.8 kcal/mol into a shallow energetic minimum on the other side. This rotational barrier can be lowered significantly if the remaining structural parameters are allowed to minimize as well during the scan. This second, relaxed scan shows that the energetic profile of the rotation is much shallower and has a barrier of less than 3.5 kcal/mol between two structures with nearly identical energies. The second minimum in the curve (structure **3a**) at 160° corresponds to a structure where one Si(CH$_3$)$_2$H unit of one C(Si(CH$_3$)$_2$H)$_3$ ligand has rotated through to the other side and the other two ligands have been left largely unchanged. This structure has been confirmed to be an energetic minimum that shows the same energy as structure **3** within numerical accuracy but is completely unsymmetrical (point group C_1). The transition structure **3b** that connects **3** and **3a** has also been identified and is calculated to be 3.4 kcal/mol higher in energy than **3** (and **3a**).

Because of the large number of internal rotations that are possible in this compound, a number of alternative structures can be imagined. For example, in the "planar" structure **1**, the noncoordinating Si−H groups are all pointing in the same direction in the horizontal mirror plane. It can be imagined that the particular Si(CH$_3$)$_2$H groups those Si−H groups belong to might be rotated by 180° in such a way that all Si−H groups point in the opposite direction of structure **1** but still coplanar with the horizontal mirror plane. This new structure (**1a**) also has C_{3h} symmetry and has been calculated to be a transition state like structure **1** but lies 3.4 kcal/mol above the previously considered structure, **1**.

Starting from structure **2**, where all three C(Si(CH$_3$)$_2$H)$_3$ ligands coordinate through one Si−H group to the yttrium atom in the center of the complex, other modes of coordination can be imagined (see Figure 3). In structure **2a**, with C_s symmetry, one of the ligands is rotated by 180° and now coordinates through two Si−H groups to the central metal atom. This increases the number of coordinating groups from three in **2** to four in **2a**. At the same time, some of the steric strain in **2** is released so that **2a** is 34.3 kcal/mol more stable than **2**. Even more steric strain can be released if not all ligands are aligned in a way where all three Si(CH$_3$)$_2$H groups are above or below the plane defined by the central yttrium atom and the three carbon atoms through which the ligands coordinate to the metal. The C(Si(CH$_3$)$_2$H)$_3$ ligands can also rotate around the Y−C bond in a way so that one Si(CH$_3$)$_2$H group (more precisely, the C−Si bond of one C−Si(CH$_3$)$_2$H group) is coplanar with the plane mentioned above. This is the case in **1**, **1a**, **3**, and **3a**, whereas in **2** and **2a**, the ligands are not in this favorable orientation. An intermediate structure **2b** has been constructed and optimized where two ligands are in the favorable "in-plane" orientation

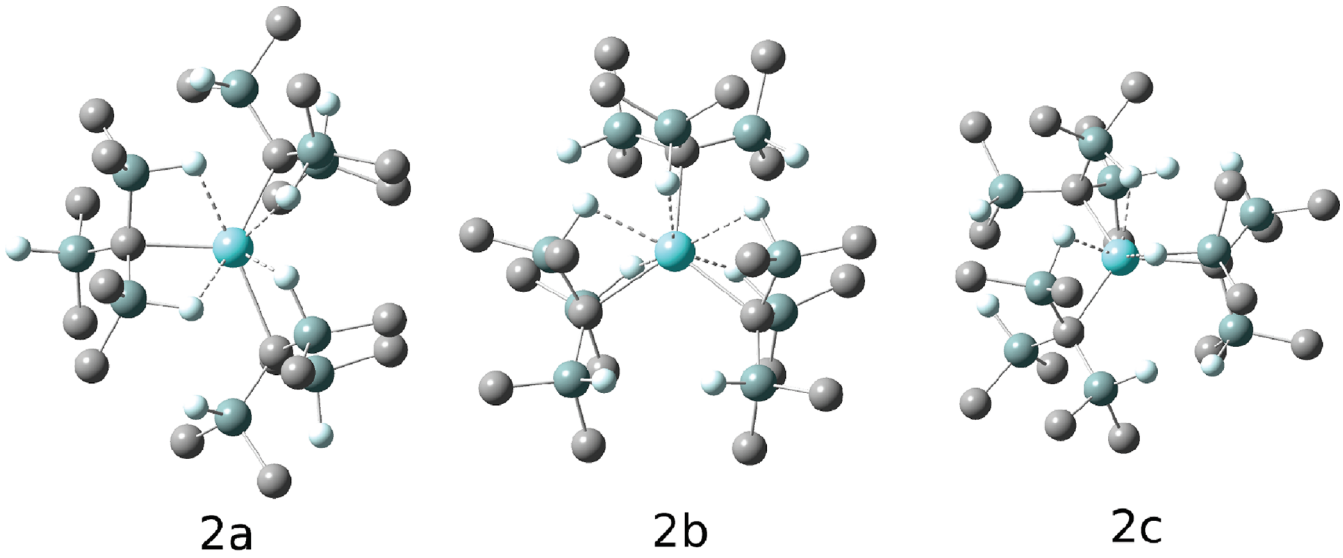


Figure 3. Structures of **2a**, **2b**, and **2c** (only SiH hydrogen atoms are shown; all other hydrogen atoms are omitted for clarity). Bonds are shown as solid lines; Si−H···Y agostic interactions are shown as dashed lines.

and hence coordinate through two Si−H bonds to the metal center each. The remaining third ligand is not in this orientation and coordinates through just one Si−H bond so that the total number of coordinating Si−H bonds in **2b** is five. With most of the steric strain now released, **2b** is only 13.3 kcal/mol less stable than **3**.

Although the comparison of **2**, **2a**, and **2b** seems to suggest that the release of steric strain is coupled to an increasing number of coordinating Si−H bonds, this is not the case. Starting from structure **2**, it is possible to create a structure that releases most steric strain but still coordinates through only three Si−H bonds like **2**. This new structure **2c** is obtained by rotating all three ligands concertedly to one side, thereby reducing the symmetry from C_{3v} to just C_3. This structure is 71.6 kcal/mol more stable than **2** and only 8.9 kcal/mol less stable than the energetic minimum **3** and, by extension, 4.4 kcal/mol more stable than **2b**, which coordinates through five Si−H bonds.

NMR Spectra. While the structural data alone demonstrate convincingly that **3** is indeed the lowest energy structure of the compound in question, it is desirable to corroborate the evidence by looking at the problem from another angle. Because of the availability of experimental ^{1}H NMR data for the system at hand, this was chosen as a basis for comparison to the computationally obtained NMR chemical shifts (see Table 1).

The Si−H protons fall into two distinct groups for all isomers. First, there are those Si−H protons that coordinate to the central metal atom, i.e., those that are in close proximity to the yttrium atom. The ^{1}H NMR signals of these protons are shifted upfield and have been characterized experimentally at 3.40 ppm. Second, there are those Si−H protons that do not coordinate and hence are further away from the central yttrium atom. Those Si−H protons give NMR signals that are shifted downfield, and they appear at 4.71 ppm in the experimental spectrum of the compound in question.

The calculated chemical shifts of structures **1** and **1a** appear to be in line with the experimental findings. The six coordinating Si−H protons are found at 3.45 ppm and 3.43 ppm, respectively. This correlates very well with the experimental finding of 3.40 ppm for those protons. Unfortunately, the agreement is undone by the calculated chemical shifts of the noncoordinating Si−H protons which are found at 5.26 ppm for **1** and 5.13 ppm for **1a**, whereas the experimentally observed chemical shifts for these protons is 4.71 ppm. A comparison to the computed NMR data of structure **2**, the pyramidalized structure favored by the initial experimental assessment, shows even worse agreement with the experimental findings. Because of the high steric strain introduced into the system by the alignment of the ligands in C_{3v}, the six equivalent Si−H protons are too far away from the central yttrium atom to coordinate effectively. Conversely, the other three Si−H protons are pushed into the proximity of the central metal atom and hence can be seen as coordinating. This reverses the experimental findings which indicate that there must be six coordinating Si−H protons and three noncoordinating protons. This clearly eliminates **2** as a possible candidate for the structure of the yttrium compound in this investigation.

Our structural candidate, **3**, is a slightly more complicated case. Instead of six equivalent coordinating Si−H protons, there are two groups of three Si−H protons that coordinate to the central metal atom but give separate signals since they are not equivalent. Those signals are predicted at 3.28 ppm and 3.70 ppm. At first glance, this seems to contradict the experimental observation of a single signal for all six coordinating Si−H protons. However, we have demonstrated above that there is only a very small barrier of 3.4 kcal/mol to the internal rotation that transforms **3** into **3a**. This internal rotation also transforms the two coordinating $Si(CH_3)_2H$ groups into each other, and hence the two calculated signals at 3.28 ppm and 3.70 ppm will merge at the temperature at which the experiment was performed to an average value of 3.49 ppm, which compares very well with the observed resonance at 3.40 ppm. Because of the extremely low barrier for the internal rotation, it will most probably not be possible to cool the sample below the coalescence point and retain

the liquid state at the same time. This quick internal rearrangement of **3** through **3b** into **3a**, or possibly from **3** to the mirror image of **3** through **1** if all ligands rotate at the same time, is also the reason why the experiment appears to indicate that all three ligands are equivalent (at all experimental temperatures) and that a mirror plane relates the two coordinating Si(CH$_3$)$_2$H groups.

Conclusions

The main findings of this investigation can be summarized as follows:

Density functional theory calculations have shown that the structure of [Y(C(SiH(CH$_3$)$_2$)$_3$)$_3$] is more complicated than first anticipated. The potential energy surface is very shallow with respect to internal rotations of the three ligands. In fact, the barriers between several low-lying minima are so small that it will be exceedingly difficult to observe the global minimum in an NMR experiment. The fast exchange of position exhibited by some Si(CH$_3$)$_2$H groups in the compound makes them appear equivalent in the experimental NMR spectrum even though our theoretical data shows that the only two possible structures containing a mirror plane which would explain the observation are transition states between the true minimum structure (**3**) or sterically unfavorable high energy structures (**2**).

The findings of the structural analysis are corroborated by the calculation of NMR chemical shifts for the Si−H protons in the compound, which show very good agreement between the experimental findings and our proposed structure of the global minimum and in turn the poor agreement of the calculated chemical shifts of the other structures in consideration of the experimental results.

Computational Details

All calculations were carried out using density functional theory (DFT) with the B3LYP density functional.[8] Correlation consistent basis sets of the Dunning type were used throughout. The yttrium atom is described by a triple-ζ-quality basis set (aug-cc-pVTZ-PP) and the accompanying effective core potential.[9] The inner shell of atoms around the yttrium metal center is described by triple-ζ-quality basis sets[10] (cc-pVTZ); this includes the carbon atoms that coordinate to the metal center, all silicon atoms, and all hydrogen atoms directly bound to a silicon atom. All other carbon and hydrogen atoms are described by a double-ζ-quality basis set[10b] (cc-pVDZ). All calculations were carried out with the Gaussian suite of programs[11] (g03).

Acknowledgment. We acknowledge the use of extensive computer time on Massey University's supercomputer facilities *DoubleHelix* and *BeSTGRID* as well as Victoria University of Wellington's supercomputer facility *Heisenberg*.

References

(1) (a) Placa, S. L.; Ibers, J. A. *Inorg. Chem.* **1965**, *4*, 778. (b) Trofimenko, S. *J. Am. Chem. Soc.* **1968**, *90*, 4754. (c) Trofimenko, S. *Inorg. Chem.* **1970**, *9*, 2493. (d) Cotton, F. A.; Jeremic, M.; Shaver, A. *Inorg. Chim. Acta* **1972**, *6*, 543. (e) Cotton, F. A.; LaCour, T.; Stanislowski, A. G. *J. Am. Chem. Soc.* **1974**, *6*, 5074. (f) Cotton, F. A.; Day, V. W. *J. Chem. Soc. Chem. Commun.* **1974**, *1974*, 415.

(2) Brookhart, M.; Green, M. L. H.; Parkin, G. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 6908.

(3) (a) Clot, E.; Eisenstein, O. *Struct. Bonding (Berlin)* **2004**, *113*, 1. (b) Scherer, W.; McGrady, G. S. *Angew. Chem., Int. Ed.* **2004**, *43*, 1782.

(4) Lein, M. *Coord. Chem. Rev.* **2009**, *253*, 625.

(5) Yan, K.; Pawlikowski, A. V.; Ebert, C.; Sadow, A. D. *Chem. Commun.* **2009**, 656.

(6) (a) Tilley, T. D.; Andersen, R. A.; Zalkin, A. *Inorg. Chem.* **1984**, *23*, 2271. (b) Rees, W. S.; Just, O.; Schumann, H.; Weimann, R. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 419. (c) Eppinger, J.; Spiegler, M.; Hieringer, W.; Herrmann, W. A.; Anwander, R. *J. Am. Chem. Soc.* **2000**, *122*, 3080. (d) Herrmann, W. A.; Eppinger, J.; Spiegler, M.; Runte, O.; Anwander, R. *Organometallics* **1997**, *16*, 1813. (e) Klimpel, M. G.; Görlitzer, H. W.; Tafipolsky, M.; Spiegler, M.; Scherer, W.; Anwander, R. *J. Organomet. Chem.* **2002**, *647*, 236.

(7) (a) Avent, A. G.; Caro, C. F.; Hitchcock, P. B.; Lappert, M. F.; Li, Z.; Wei, X.-H. *Dalton Trans.* **2004**, 1567. (b) Barker, G. K.; Lappert, M. F. *J. Organomet. Chem.* **1974**, *76*, C45. (c) Schaverien, C. J.; Orpen, A. G. *Inorg. Chem.* **1991**, *30*, 4968.

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

(9) Peterson, K. A.; Figgen, D.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2007**, *126*, 124101.

(10) (a) Woon, D. E.; Dunning, J. T. H. *J. Chem. Phys.* **1993**, *98*, 1358. (b) Dunning, J. T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(11) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.

CT100663Y

JCTC Journal of Chemical Theory and Computation

Protein Free Energy Corrections in ONIOM QM:MM Modeling: A Case Study for Isopenicillin N Synthase (IPNS)

Tsutomu Kawatsu,[†] Marcus Lundberg,[†] and Keiji Morokuma*[,†,‡]

Fukui Institute for Fundamental Chemistry, Kyoto University, 34-4 Takano Nishihiraki-cho, Sakyo-ku, Kyoto 606-8103, Japan, and Cherry L. Emerson Center for Scientific Computation and Department of Chemistry, Emory University, Atlanta, Georgia 30322, United States

Received September 27, 2010

Abstract: The protein environment can have significant effects on the enzyme catalysis even though the reaction occurs locally at the reaction center. In this paper, we describe an efficient scheme that includes a classical molecular dynamics (MD) free-energy perturbation (FEP) correction to the reaction energy diagram, as a complement to the protein effect obtained from static ONIOM (QM:MM) calculations. The method is applied to eight different reaction steps, from the O_2-bound reactant to formation of a high-valent ferryl-oxo intermediate, in the nonheme iron enzyme isopenicillin N synthase (IPNS), for which the QM:MM energy diagram has previously been computed [Lundberg, M. et al. *J. Chem. Theory Comput.* **2009**, *5*, 220−234]. This large span of the reaction coordinate is covered by dividing each reaction step into microsteps using a virtual reaction coordinate, thus only requiring ONIOM information about the stationary points themselves. Protein effects are important for C−H bond activation and heterolytic O−O bond cleavage because both these two steps involve charge transfer, and compared to a static QM:MM energies, the dynamics of the protein environment changes the barrier for O−O bond cleavage by several kcal/mol. The origin of the dynamical contribution is analyzed in two terms, the geometrical effect caused by the change in average protein geometry (compared to the optimized geometry) in the room temperature MD simulation with the solvent, and the statistical (entropic) effect resulting from fluctuations in the interactions between the active site and the protein environment. These two effects give significant contributions in different steps of the reaction.

1. Introduction

When modeling enzymatic reactions, it is common to separate the reactivity of the active site from the effects of the surrounding protein matrix. This approximation seems especially valid for transition metal enzymes, for which the activity of biomimetic complexes[1] indicates that the reactivity mainly depends on the electronic structure of the metal center.[2] However, enzymes with very similar active sites catalyze different reactions, and an explicit description of the protein environment is necessary to fully understand the reaction mechanism, relative reaction rates and substrate selectivity.

QM/MM (quantum mechanics/molecular mechanics) models take advantage of the separation between reaction active site and protein environment by treating these regions at different levels of theory.[3] Our group has developed the ONIOM multiscale method that calculates the total energy of the molecular system by an extrapolation scheme including different QM and MM calculations.[4−9] The QM:MM label

* To whom correspondence should be addressed: E-mail: morokuma@fukui.kyoto-u.ac.jp. Telephone: +81-75-711-7843.

[†] Kyoto University.

[‡] Emory University.

10.1021/ct1005592
Published on Web 12/29/2010

separates ONIOM from standard QM/MM methods that employ additive schemes. The interaction between the QM part and the MM part can be included in the calculations either classically by mechanical embedding (ONIOM-ME) or semiclassically by electronic embedding (ONIOM-EE).[9] We have previously used the ONIOM QM:MM method to describe the protein effects on several metalloenzyme reactions.[10] Recently, we have employed an advanced algorithm, the "fully coupled macro/micro-iterative" optimization scheme,[11] to efficiently locate transition states in complex molecular systems, specifically in mammalian glutathione peroxidase,[12] isopenicillin N synthase (IPNS),[13] and methylmalonyl-CoA mutase.[14,15] The protein environment influences the description of the reactivity, but the effect on the calculated energy barriers varies significantly between different enzymes, significantly lowering the barrier in methylmalonyl-CoA mutase, while having only a modest effect in glutathione peroxidase.

However, describing a large system with optimization techniques requires special care to avoid artificial changes in geometry that can lead to large errors in relative energies.[9] Static methods also cannot describe situations where the environment changes during the chemical reaction, for example, new alignment of side chains or solvent water, thermal fluctuations, or large-scale protein motions. Because of the lack of geometric polarization, the static approach may overestimate electrostatic effects. In the present study, we replace the static interactions between protein and QM region by classical free-energy corrections from dynamical sampling of millions of protein configurations. QM/MM approaches with free-energy perturbation (FEP) have previously been used to describe reactions in both protein and solvent.[16−21] To separate the present approach from others efforts in the area, we use the description QM:[MM-FEP] for the ONIOM QM:MM approach, where the effect of the MM layer is described by free-energy perturbation.[22]

One of the main objectives of the present method is the capability to estimate the dynamic effects on the reaction energy profile of complicated enzymatic reactions, for example, a multistep redox reaction in a transition metal enzyme. The difference in electronic structure and nuclear coordinates between two stationary states can be large, so in FEP each reaction step is divided into several intermediate points, for example, by following selected reaction coordinates (typically bond distances) or the intrinsic reaction coordinate (IRC). However, transition-metal systems have complicated multidimensional reaction coordinates, so to avoid a detailed mapping, we adopt the standard alchemical FEP technique that only requires information about the initial and the final state.[23] Intermediate points are generated by a virtual reaction coordinate that gradually mixes the initial and the final state. The required information about these states is obtained by full QM:MM optimizations of all stationary points, including transition states, using the fully coupled Hessian algorithm.[9]

Transition-metal systems require relatively expensive QM methods, for example, hybrid DFT. We therefore freeze the geometry of the QM part and perform the FEP calculations with fully classical samplings (ONIOM-ME). In this approximation the QM energy (for a given QM geometry) does not depend on the protein structure, so only one calculation of the QM wave function is required for each geometry. This approximation is similar to the QM/MM-FE method,[24,25] and calculations by Rod et al.,[18,26] show that this method differs by less than 3 kcal/mol from their more elaborate QTCP method. In this context, it must be kept in mind that modeling of transition metal reactions is a difficult task and that the inherent error in the QM treatment has been estimated to be 3−5 kcal/mol.[2]

We apply the QM:MM method with free-energy corrections to the nonheme iron enzyme isopenicillin N synthase that catalyzes the formation of isopenicillin N (IPN), a key reaction in penicillin synthesis that is still used in large-scale production.[27] Our previous QM and QM:MM studies[13,28,29] identified 19 intermediates and transition states for the reaction leading from the ACV substrate to the IPN product. For the free-energy treatment, we selected the first half of this reaction, C−H bond activation from the iron-bound dioxygen species, followed by heterolytic O−O bond cleavage to form a ferryl-oxo species (9 stationary points). Of the two alternative mechanisms for Fe(IV)−oxo formation previously investigated, only the "ligand donor" mechanism is chosen here, partly because it shows larger protein effects.

In the following sections, we first describe the computational details of the QM:[MM-FEP] method, followed by a presentation of the IPNS free energy diagram and a discussion of how it differs from a potential energy diagram calculated with the standard QM:MM optimization method.

2. Methods of Computation

2.1. Free Energy in the ONIOM QM:MM Scheme. The relative energy for a standard ONIOM QM:MM calculation is obtained as follows:[6]

$$\Delta E_{\mathrm{ONIOM}} = \Delta E_{\mathrm{QM}}^{\mathrm{model}} + \Delta E_{\mathrm{MM}}^{\mathrm{real}} - \Delta E_{\mathrm{MM}}^{\mathrm{model}} \qquad (1)$$

where real includes all atoms in the system and model includes the selected reaction center, with hydrogen link atoms for truncated covalent bonds. In mechanical embedding (ME), the interaction between model and real system is described at the low level of theory of the real system. The QM:MM-ME energy thus includes a QM-level description of the relative energy, and an MM-level description of the protein effect on the relative energy. An alternative to mechanical embedding is electronic embedding (EE), where the QM-MM interactions are evaluated semiclassically by including the MM point charges in the QM calculation of the model system. For a free-energy calculation, the EE approach becomes very expensive as it requires a new QM calculation for each position of the surrounding atoms. We take advantage of the speed of the ME approximation and perform fully classical free-energy calculations.

In our QM:[MM-FEP] approximation, the free energy difference between two states is calculated as

$$\Delta F_{\mathrm{ONIOM}} = \Delta E_{\mathrm{ONIOM}} + \Delta f_{\mathrm{QM}}^{\mathrm{model}} + \Delta F_{\mathrm{FEP}}^{\mathrm{int/NB}} - \Delta E_{\mathrm{MM}}^{\mathrm{int/NB}} \qquad (2)$$

where Δf_{QM}^{model} is the free energy correction of the model system obtained from a Hessian calculation using the harmonic oscillator approximation. In the IPNS QM:MM potential diagram this term has been determined from calculations on an active-site QM-only model.[28] $\Delta F_{FEP}^{int/NB}$ and $\Delta E_{MM}^{int/NB}$ are the classical nonbonded interactions between the QM model system and the protein environment, evaluated using the FEP method and the static ONIOM-ME approach, respectively. By the use of the third and fourth terms, we replace the nonbonded interactions in standard ONIOM by the free energy of the model system and additionally solvent effect.

Note that the free-energy contributions in QM:[MM-FEP] does not explicitly include the protein−protein interactions. The total free energy of the protein fluctuates on a scale much larger than the energy differences between two stationary points, and it is challenging to converge the total energy. These interactions are still taken into account in the calculations, because the low-energy protein configurations dominate the calculation of nonbonded interactions between the model system center and the protein environment.

Free-energy calculations are performed with the real system at the low (MM) level with the coordinates of the model system frozen during the simulations. We have neglected the difference between terms including the hydrogen link atoms in Δf_{QM}^{model} compared to the free energy of the original covalent bonds that were truncated to form the model system. This effect is difficult to estimate correctly in the FEP approach,[30] but the effect is not likely to change significantly during the reaction and is therefore neglected. We have also neglected the cross term between the free energy of the QM model system and the protein environment, that is, how the protein environment affects the QM vibrations and vice versa.

2.2. Free-Energy Perturbation Method. In the alchemical FEP technique, an arbitrary number of intermediate points can be created by increasingly mixing two stationary points using a dual-topology method.[31] Here, we use λ_i, a virtual reaction coordinate that changes from 0 to 1 with increasing i. We run the molecular dynamics (MD) simulation between X (initial) and Y (final) states along the path of virtual intermediate states. The MD Hamiltonian is

$$H_{MD}^{i}(X,Y;R) = H_{env}(R) + (1-\lambda_i)H_X(R) + \lambda_i H_Y(R) \quad (3)$$

where $H_{env}(R)$ is the Hamiltonian excluding the QM model system, as well as the interactions between the model system and protein environment. $H_X(R)$ and $H_Y(R)$ are Hamiltonians of the model system in the two stationary states, X and Y, including their interaction terms with the protein environment. $R \equiv R(X,Y;\lambda_i)$ are the atomic coordinates of the protein environment, obtained by the molecular dynamics using the Hamiltonian $H_{MD}^{i}(X,Y;R)$. As the set of $(X,Y;\lambda_i)$ defines the Hamiltonian of a mixed chemical state, only information about initial and final states are required.

For the calculations of free energy, we define H_X^{NB} and H_Y^{NB} that exclude the bonded interaction terms (for link atoms) between the model system (redox center) and the protein environment from H_X and H_Y. The change in free energy $\Delta F_{FEP}^{int/NB}$ is the sum of free energy differences along the virtual reaction path

$$\Delta F_{FEP}^{int/NB}(X,Y) = \sum_{i=1}^{n} \Delta F_{i,i+1}(X,Y) \quad (4)$$

where the free energy difference for each virtual step is calculated as

$$\Delta F_{i,i+1}(X,Y) = -k_B T \ln\left\langle \exp\left\{ \frac{H_X^{NB}[R(X,Y;\lambda_i)] - H_Y^{NB}[R(X,Y;\lambda_i)]}{k_B T}(\lambda_{i+1}-\lambda_i) \right\} \right\rangle_i \quad (5)$$

Here, k_B is the Boltzmann constant, and T is the temperature. Where we used the energy change in the reaction coordinate from λ_i to λ_{i+1}, $H_{MD}^{i+1} - H_{MD}^{i} = -(H_X - H_Y)(\lambda_{i+1} - \lambda_i) \approx (H_X^{NB} - H_Y^{NB})(\lambda_{i+1} - \lambda_i)$.

From the FEP calculations we obtain the part of the free energy caused by the dynamics of the protein environment (so-called dynamical contribution, thereafter) $\Delta F_{FEP}^{int/NB}(X,Y) - \Delta E_{MM}^{int/NB}(X,Y)$ for each pair of stationary points, X and Y. For convenience we hereafter use the notations $\Delta F_{XY}^{FEB} \equiv \Delta F_{FEP}^{int/NB}(X,Y)$ and $\Delta E_{XY}^{opt} \equiv \Delta E_{MM}^{int/NB}(X,Y)$.

2.3. Geometrical and Statistical Effects. To better understand the origin of the QM:[MM-FEP] dynamical effects on the reaction energy diagram, we separate the contributions to the free energy into two parts: the *geometrical effect* and the *statistical effect*. The geometrical effect is the result of a change in average protein geometry in the room temperature MD simulation compared to the optimized structure. The statistical effect comes from fluctuations around the average geometry, because favorable protein geometries, that is, those that represent low-energy pathways, give larger contributions in the calculation of the free energy for a reaction step. ΔF_{XY}^{FEB} is the difference of the free energy interaction between state X and Y and includes all dynamical effects, while ΔE_{XY}^{opt} is the difference in protein interaction between optimized geometries, and excludes all dynamical effects. Here, we define a value that includes the geometrical effect, but, not the statistical effect. For this purpose we investigated the QM-MM interaction difference between $\lambda = 0$ (state X) and 1 (state Y) in MD simulations with the state-X geometry and charges. The interaction energy changes as

$$\Delta H_{XY}(X) \equiv H_X^{NB}(X) - H_Y^{NB}(X) \quad (6)$$

where $H_X^{NB}(X) \equiv H_X^{NB}(R(X,Y,0)) = H_X^{NB}(R(Y,X,1))$ MM Hamiltonians of the nonbonded interaction between the QM part and protein environment in the MD simulation were defined in section 2.2. $H_Y^{NB}(X) \equiv H_Y^{NB}(R(X,Y,0)) = H_Y^{NB}(R(Y,X,1))$ is written in same indices, but the interaction is between the state-Y QM part and the state-X protein environments. Because state-X protein geometry is not the optimized static geometry, $\Delta H_{XY}(X)$ is different from ΔE_{XY}^{opt}. The difference is the geometrical effect. $\Delta H_{XY}(X)$ includes the geometrical effect for each geometry, but not the statistical effect. The average of ΔH_{XY} then includes the average geometrical effect for the reaction barrier between states X and Y. We define the average as

$$\langle \Delta H \rangle_{XY} \equiv \langle [\Delta H_{XY}(X) - \Delta H_{YX}(Y)]/2 \rangle \quad (7)$$

We use the values $\langle \Delta H \rangle_{XY} - \Delta E_{XY}^{\mathrm{opt}}$ to estimate the geometrical contribution and the remaining free-energy contribution, $\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}$ is the statistical effect. We must notice that because $\langle \Delta H \rangle_{XY}$ includes only the situation of $\lambda = 0$ and 1, the comparison between $\langle \Delta H \rangle_{XY}$ and $\Delta F_{XY}^{\mathrm{FEP}}$ is inconsistent. However, when states X and Y are enough close, $\langle \Delta H \rangle_{XY}$ is reasonable approximation of the average potential energy shift from state X to Y that excludes the statistical effect.

For discussing the details, we partition the geometrical effect into contributions from each single residue i, $\Delta\Delta H_{XY}^{i}(X) \equiv \Delta H_{XY}^{i}(X) - \Delta E_{XY}^{\mathrm{opt},i}$ for an average geometry of state-X MD simulation, where $\Delta H_{XY}^{i}(X)$ and $\Delta E_{XY}^{\mathrm{opt},i}$ are the contributions of residue i to $\Delta H_{XY}(X)$ and $\Delta E_{XY}^{\mathrm{opt}}$, respectively.

The fluctuation of $\Delta H_{XY}(X)$ can correlate to the statistical effect. To investigate the fluctuation, we define the standard deviation s_{XY} as

$$s_{XY} \equiv \sqrt{\langle [\Delta H_{XY}(X)]^2 \rangle - \langle \Delta H_{XY}(X) \rangle^2} \quad (8)$$

We also define s_{YX} for the backward reaction from state Y to state X in the state-Y protein environment. We calculate the standard deviation (s_{XY}^{i}) of the single residue contribution to the interaction change ($\Delta H_{XY}^{i}(X)$) as

$$s_{XY}^{i} \equiv \sqrt{\langle [\Delta H_{XY}^{i}(X)]^2 \rangle - \langle \Delta H_{XY}^{i}(X) \rangle^2} \quad (9)$$

2.4. Computational Model. The protein setup and the ONIOM system are described in detail in reference.[13] Here we use the small 65-atom model system including Fe, a water ligand, selected parts of the three amino acids His214, Asp216, His270, and the reactive part of the substrate, see Figure 1. The small size of the model system increases the protein effects and makes it easier to evaluate the difference of the static and dynamic approaches. QM calculations were performed with the density functional B3LYP. The 6-31G(d) basis set was used for the geometry optimizations and Hessian calculations, while 6-311+G(d,p) was used for energy evaluations.

The classical nonbonded interactions between model and real system depend on the van der Waals parameters and the assigned point charges. Atoms outside the model system were assigned parameters from the Amber94 force field[34] to be able to compare with the previous QM:MM calculations.[13] Atoms in the model system are assigned point charges from RESP[35] calculations of the model system for each stationary point, using the Gaussian[36] standard geometry for the ESP calculations and the Antechamber module of Amber[37] for the first step of the RESP fitting, see Supporting Information. Charges for the part of the substrate that is not included in the model system were assigned from a calculation in the reactant state and were not changed during the reaction.

2.5. Simulation Details. We started the free-energy calculations based on the optimized QM:MM geometries of the stationary states as reported in ref 13. The protein was placed in an approximately 80 × 68 × 57 Å^3 water box including ~7700 TIP3P water molecules and 11 sodium ions

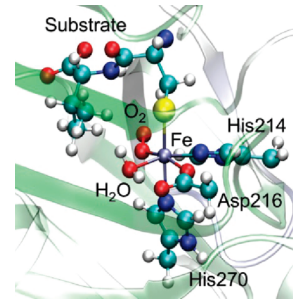


Figure 1. QM:MM model system of isopenicillin N synthase (QM atoms in ball and stick). All protein illustrations in the present paper are prepared using the VMD program.[32,33]

with periodic boundary condition. Simulations were run using the NAMD molecular dynamics program.[38,39] Five thousand steps of energy minimization and 1.5 ns equilibration were applied at each stationary point before the start of the FEP calculation. All simulations used a 1 fs time step. The temperature was controlled at 298 K using a Langevin thermostat every 5 fs.

We split each reaction step into 24 virtual steps (25 states) using the virtual λ coordinate ($\lambda_i = 0$, 0.001, 0.01, 0.05, 0.1, ..., 0.9, 0.95, 0.99, 0.999, 1.0). We use an unevenly spaced reaction coordinate to reduce the impact of steric clashes when the final state is turned on. There are nine stationary points along the reaction path, and the total number of intermediate states is 193. Each virtual step includes 100 ps equilibrium followed by 500 ps sampling simulations. Molecular dynamics is performed with the Hamiltonian for state i, $H_{\mathrm{MD}}^{i}(X,Y;R)$, and the difference in nonbonded interaction energies between two topologies (i and $i + 1$) is evaluated each 10 fs.

Average values of the exponential and statistical errors were estimated using the bootstrap method.[40,41] Our program uses a subroutine for the inverse error function calculation in Ooura's Mathematical Software package.[42] For each reaction step, we ran simulations both in forward ($\lambda = 0$ to 1) and backward ($\lambda = 1$ to 0) directions. The final result is the average of these two values, and the error bars are determined by the difference of the two simulations. The method presents statistically correct estimation of an average value in biased sampling.

3. Results and Analysis

3.1. QM:MM-ME and -EE Potential Energy Profiles. The use of a fully classical mechanical embedding (QM:MM-ME) potential is critical in our method as it allows for longer sampling times and better convergence of the FEP calculations. To check how the classical approximation of the protein−core interactions matches the semiclassical electronic embedding approach, we calculated the static energy diagram using both methods, see Figure 2. All comparisons are made with the same protein geometry for each stationary point, and the energies in Figure 2, therefore, represent situations without any geometric relaxation of either core or protein. This assumption overestimates the electrostatic effects and the difference between the methods, but still serves a purpose for a general comparison. The present

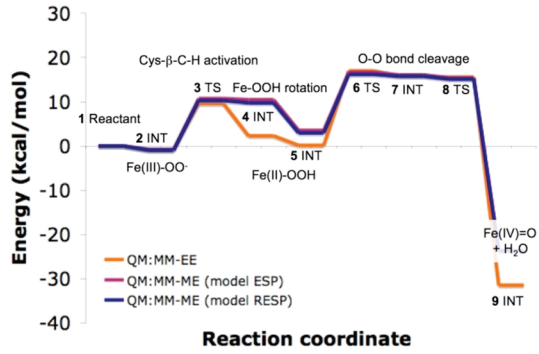


Figure 2. Comparison between ONIOM-EE and ONIOM-ME with charges (ESP and RESP) of the model system updated at every stationary point.

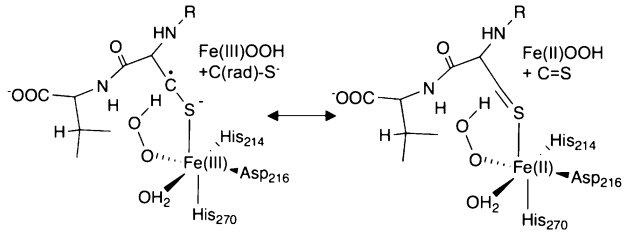


Figure 3. Resonance structures for stationary point **4**. The difference between the left and the right structure is an electron transfer from substrate to iron.

ONIOM-ME energy diagram is different from the energy diagram in ref 13. In the present study, the model charges are updated at each stationary point, but in the previous study the MM parameters were constant to provide a continuous energy surface for optimization of a large number of stationary points, including transition states. To facilitate the discussion in this section and later, each stationary point is described by a number in bold, representing the order it appeared along the reaction energy diagram. For details see ref 28.

Counting all stationary points, the mean absolute deviation between ME and EE approximations is 2.2 kcal. Major deviations between the two methods appear for stationary points **4** (iron-bound peroxide) and **9** (ferryl-oxo + water). In the first case, **4**, two alternative electronic structures can be drawn, see Figure 3, with the difference being a charge transfer between substrate and iron. This charge transfer leads to significant electrostatic repulsion from the surrounding protein, and while the ME calculation indicates a complete charge transfer (no spin population on the substrate), in the EE calculation there remains some unpaired spin population on the substrate carbon (−0.14). The difference in electron density between ME and EE can be ascribed as a polarization effect of the surrounding protein. In the second case, **9**, heterolytic O−O bond cleavage leads to release of water from the active site, and new direct interactions between QM and MM atoms. Fortunately these effects are not essential for the modeling of a reaction mechanism because product formation is exothermal and the energy of reaction does not affect the barriers of the proceeding steps of the reaction.

If we only compare the barrier heights of the three transition states, from state **2** to **3**, **5** to **6**, and **7** to **8**, the mean absolute deviation is only 0.67 kcal/mol. Although the classical approximation shows differences compared to the semiclassical approximation for certain steps, the potential energy diagrams for the present reaction are mostly reasonable. We therefore employ the classical charges for our FEP calculations.

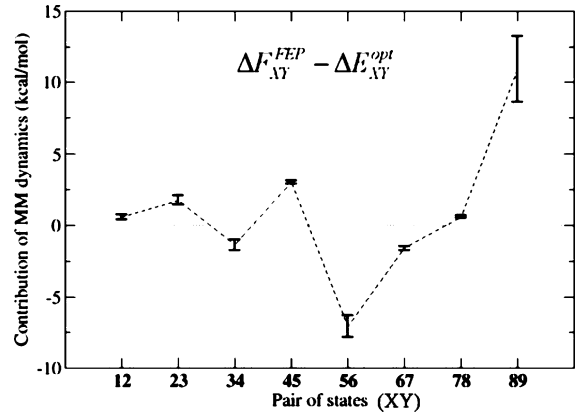


Figure 4. Contribution of MM dynamics, relative to a static QM:MM description, on the relative energies of neighboring stationary points *X* and *Y*.

3.2. Dynamical Contribution to the Free-Energy Diagram. The dynamical contributions to the free-energy diagram $\Delta F_{XY}^{\mathrm{FEP}} - \Delta E_{XY}^{\mathrm{opt}}$ are shown in Figure 4. In this figure each reaction step is written as a pair of initial (*X*) and final state (*Y*). Effects exceed 3 kcal/mol for pairs $XY =$ **45**, **56**, and **89** and must, therefore, be considered significant. The transition from state **4** to state **5** is a rotation of the peroxide formed after C−H bond activation, a step, where there is a large change in the active site geometry. The transition from **5** to **6** corresponds to an electron transfer from iron to an antibonding π^* O−O orbital, a step with large electrostatic effects from the protein environment, see discussion in ref 13. The transition from **8** to **9** is the release of H_2O from the reaction center after O−O bond cleavage, and the QM water molecule makes several new hydrogen bonds with MM residues and water molecules. The error bars for the FEP calculations are relatively small, with the exception of pair **89** where the released water molecule sometimes has strong steric interactions with the MM atoms as the virtual coordinate λ changes from 0 to 1.

The calculated free-energy diagram is shown in Figure 5. The blue line is the static QM:MM result (with updated RESP charges) and the red line is the QM:[MM-FEP] result. The QM:[MM-FEP] results represent a frozen core and a flexible protein, and to make a relevant comparison, the ONIOM-ME results are obtained with the same core geometry and with the protein reoptimized after the RESP charges had been updated.

Compared to the static ONIOM-ME description, the free-energy description leads to an increase in the C−H activation barrier (state **2** to **3**) and a decrease in the barrier for O−O bond activation by significantly stabilizing state **6** relatively to **5**. For evaluation of the reaction mechanism, the most significant difference is the predicted rate-limiting step. The ONIOM-ME (RESP) diagram suggests that O−O bond activation is rate-limiting. The free-energy approach gives a higher barrier for C−H bond activation step (**2** to **3**)

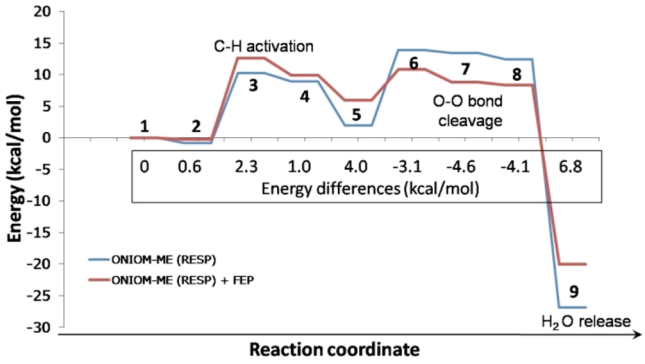


Figure 5. Free-energy diagram for the formation of an Fe(IV)-oxo intermediate after dioxygen binding. The results of the FEP approach are compared to results from a static approach where the protein has been optimized (ONIOM-ME). The relative energy differences between the two approaches are shown in the center box.

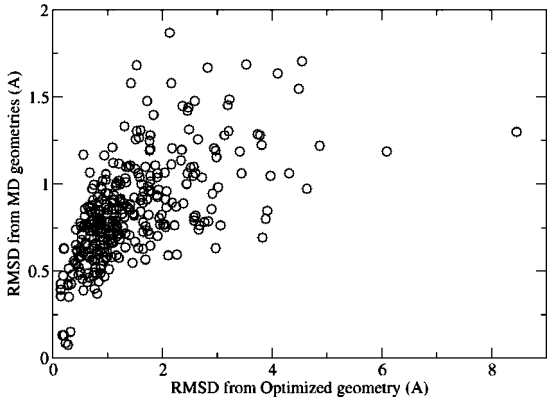


Figure 6. rmsd of the geometry of heavy atoms on each residue. The horizontal axis indicates rmsd between the optimized geometry and the MD averaged geometry. The vertical axis indicates rmsd between the MD averaged geometry and geometries in 10 MD snapshots, from $XY =$ **34** forward simulation. Each point represents a protein residue.

compared to O−O bond cleavage (**5** to **6**), which is consistent with kinetic isotope experiments that show that C−H activation is at least partly rate-limiting.[43] However, the two barriers are relatively close in both cases and the uncertainties in both QM and MM treatment makes it difficult to use the relative barriers as a reliable benchmark. The computed free-energy of C−H activation barrier is 12.9 (±0.3) kcal/mol, where the error bars reflect the statistical error of the FEP calculation. The value is significantly lower than the experimentally estimated reaction barrier of 16.8 kcal/mol,[44] but that barrier is based on a DFT calculation with the B3LYP functional. This method may underestimate barriers of simple hydrogen atom transfer reactions.[45,46]

We discuss the geometrical and statistical effects in the next two subsections.

3.3. Statistical vs Geometrical Effects in the Free Energy Diagram. The geometrical effect is caused by the change in average protein geometry compared to the optimized structure. The statistical effect comes from fluctuations around the average geometry. In Figure 6, each protein residue is represented by a circle. The horizontal axis indicates the rmsd value between the MD average geometry

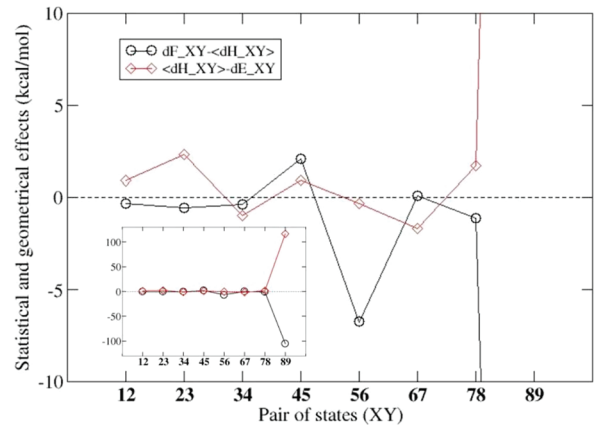


Figure 7. Comparison between the statistical contribution, $\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}$ (black circle) and the geometrical contribution $\langle \Delta H \rangle_{XY} - \Delta E_{XY}^{\mathrm{opt}}$ (red diamond) of the protein dynamics on the free energy profile.

and the optimized geometry, that is, how much a residue moves on average in MD simulations. The vertical axis indicates the rmsd in geometry of 10 MD snapshots from the average MD geometry. This value correlates to the flexibility of the residues, that is, how much their positions fluctuate from their average MD positions. Many of these flexible residues are on the protein surface, which is reasonable because the MD system includes explicit water molecules that allow surface residues to move during the simulation. However, we also note that many residues inside the protein also show significant flexibility. We found residues for which both RMSDs are large, and these geometry changes may represent the geometrical and statistical effects. For a more detailed analysis of geometry changes during MD simulations, see the Supporting Information.

Here, we try to determine which effect controls the protein effect on each barrier of the entire reaction free energy profile. We calculated three type of nonbonded interaction energy, $\Delta F_{XY}^{\mathrm{FEP}}$, $\Delta E_{XY}^{\mathrm{opt}}$, and $\langle \Delta H \rangle_{XY}$ that include both geometrical and statistical effects, no effect and the geometrical effect, respectively. When $\Delta F_{XY}^{\mathrm{FEP}}$ and $\Delta E_{XY}^{\mathrm{opt}}$ have similar values, there is no net dynamical contribution to the free energy profile. $\langle \Delta H \rangle_{XY} - \Delta E_{XY}^{\mathrm{opt}}$ and $\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}$ are the geometrical and statistical contributions, respectively. The results for $\Delta F_{XY}^{FEP} - \langle \Delta H \rangle_{XY}$ and $\langle \Delta H \rangle_{XY} - \Delta E_{XY}^{\mathrm{opt}}$ for all reaction steps are shown in Figure 7.

We found three cases: (a) the geometrical effect $|\langle \Delta H \rangle_{XY} - \Delta E_{XY}^{\mathrm{opt}}|$ is larger than the statistical effect ($XY =$ **12**, **23**, **34**, **67**, and **78**); (b) the statistical effect $|\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}|$ is larger ($XY =$ **45** and **56**); and (c) both values are very large ($XY =$ **89**). In the first case, the geometrical effect dominates the dynamical contribution of the free energy profile. In the free energy diagram (Figure 5), this effect increases the transition state barrier for C−H bond activation from **2** to **3** and affects the energy of steps **34** and**67**. For $XY =$ **12** and **78**, geometrical and statistical effects partly cancel and the total dynamical contribution is small. In second case, the statistical effect dominates the dynamical contribution. This effect decreases the exothermicity of $XY =$ **45** (peroxide rotation) and decreases the reaction barrier of O−O bond activation ($XY =$ **56**). In the third case ($XY =$

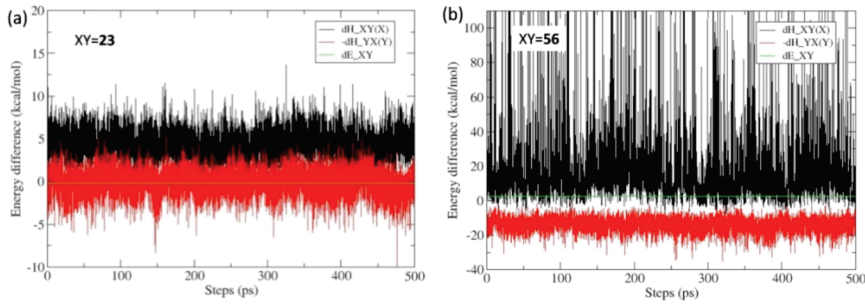


Figure 8. Fluctuation of the interaction energy change $\Delta H_{XY}(X)$ (black line) and the negative reverse $-\Delta H_{YX}(Y)$ (red line) in FEP calculations. (a) $XY = \mathbf{23}$ and (b) $XY = \mathbf{56}$. Green line is the interaction energy difference $\Delta E_{XY}^{\mathrm{opt}}$ between state X and Y in the optimized geometries.

89), both effects are large, but partly cancel, and as the geometrical effect is larger there is a net positive contribution to the total dynamical effect, see Figure 4. However, the case $XY = \mathbf{89}$ may contain an artificial effect of broken criterion of the perturbation calculation (see section 4).

To clarify how these energetic results are coupled to the molecular dynamics simulation, we show the fluctuation of ΔH_{XY} in two typical cases, one where the geometric effect dominates ($XY = \mathbf{23}$) and one where the statistical effect dominates ($XY = \mathbf{56}$), see Figure 8. Black and red lines indicate the forward direction $\Delta H_{XY}(X)$ and the negative reverse direction $-\Delta H_{YX}(Y)$. The green straight line is $\Delta E_{XY}^{\mathrm{opt}}$. In the case of $XY = \mathbf{23}$, Figure 8a, the fluctuations of both black and red lines are small. That is, most geometries have similar opportunity for the reaction from state X to Y or vice versa. All these geometries contribute similarly to the free energy profile and the statistical effect is small. However, the green line, representing the static calculation, is shifted relative to the center of the black and red lines that represent the dynamical calculation. Looking back at the energy contributions in Figure 7, it is clear that the geometrical effect, rather than the statistical effect, dominates the protein contribution to the free energy profile.

In the case of $XY = \mathbf{56}$, Figure 8b, the fluctuation of the black line is large. The possibility that the state X switches to state Y at the geometry R is proportional to $\exp(-\Delta H_{XY}(R)/k_{\mathrm{B}}T)$. Therefore, geometries with lower ΔH_{XY} have larger opportunity to react from state X to Y. When the fluctuation of ΔH_{XY} is large, the reaction is dominated by the few geometries with very low values of ΔH_{XY} rather than the geometric average, and the statistical effect becomes large. Other geometries do not contribute to the reaction from state X to Y.

When the black and red lines cross, these conformational pairs in the same energy can switch states XY and its protein geometries without extra energy. Such condition often appears at the top of the free energy potential barrier. $XY = \mathbf{56}$ has less crossing possibility than $XY = \mathbf{23}$ and intermediate states of X and Y must contribute to the reaction instead of pure states of X and Y. We notice that these two simulations are independent and time axis can shift.

As a measure of the stability of the calculated QM-MM interactions, we use the standard deviation s_{XY} for the interaction energy change ΔH_{XY}, see section 2.3. s_{XY} is defined for the forward reaction from state X to state Y in

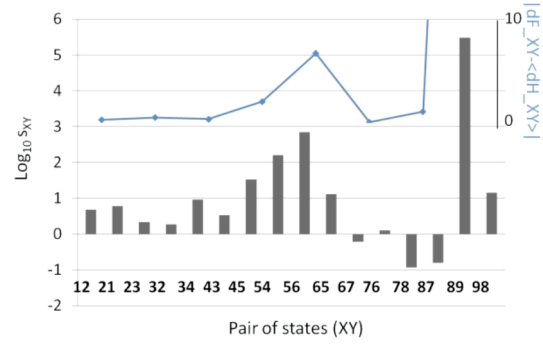


Figure 9. Fluctuation of the interaction change $\Delta H_{XY}(X)$ from state X to state Y for all steps, comparing with pairs of states of stationary points for the IPNS enzymatic reaction, see Figure 5. The vertical axis is the log scale of the standard deviations s_{XY} of the interaction shift $\Delta H_{XY}(X)$. s_{XY} is in kcal/mol. The blue line is the absolute values of the statistical effects to compare with standard deviations.

the state-X protein environment in eq 7. The protein dynamics presents various barrier heights of the reaction steps in time. And when the barrier is low, the reaction can occur easily. During the molecular dynamics simulation, the reaction is more likely to take place through the geometry that has lower energy difference ΔH_{XY} between two stationary points, and the free energy of the reaction ΔF_{XY} becomes lower than in the average geometry. The situation would be the same the reverse reaction with s_{YX} and $\Delta F_{YX} \equiv -\Delta F_{XY}$. When s_{XY} and s_{YX} are similar, the net effect is zero. However, when one of them is larger than the other, fluctuations give a net effect on ΔF_{XY} (if s_{XY} is larger than s_{YX}, ΔF_{XY} decreases). Figure 9 shows s_{XY} and s_{YX} for the eight reaction steps. For example, $s_{\mathbf{54}}$ is much larger than $s_{\mathbf{45}}$for the opposite direction, and $s_{\mathbf{56}}$ is much larger than $s_{\mathbf{65}}$. $\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}$ values of $XY = \mathbf{45}$ and **56** are largely positive and negative, respectively, see Figure 7. These dynamical contributions mainly originate from the statistical or entropic effect of the thermal fluctuations. We put the graph of $|\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}|$ to compare with the log of standard deviations. Good correlation is shown in Figure 9. For sign of $\Delta F_{XY}^{\mathrm{FEP}} - \langle \Delta H \rangle_{XY}$, see Figure 7.

3.4. Residue Contributions of the Geometrical and Statistical Effects. Figure 10a−d show computed $\Delta\Delta H_{XY}^{i}(X)$ for four pairs with significant dynamical effects, see Figure 4, $XY = \mathbf{23}$, **67** (weak statistical effect) and $XY = \mathbf{45}$ and **56** (strong statistical effect), respectively.

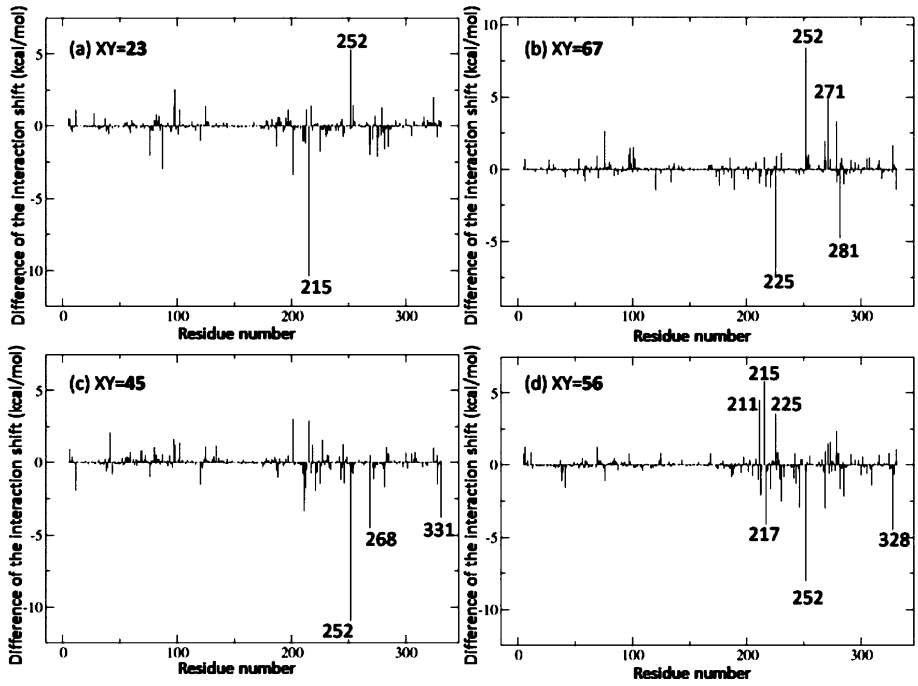


Figure 10. Differences of the interaction energy change $\Delta\Delta H^{i}_{XY}(X)$ between the average MD stationary point geometry and ONIOM QM:MM optimized geometry for individual residue *i*, namely, contributions of residue *i* to the geometrical effect of protein dynamics on the free energy profile. (a, b) Cases of the weak statistical effect ($XY =$ **23** and **67**). (c, d) Cases of the strong statistical effect ($XY =$ **45** and **56**).

In each case, several residues contribute significantly, resulting in partial cancellation of the geometrical effect. A common significant residue, Asn252, has a hydrogen bond with the residue Asp216; Asn252 can rotate almost 180 degrees between the stationary points in MD simulation. This residue is located close to the ligand water molecule that is involved in the reaction. The energetic effect of Asn252 is large in both steps **56** and **67**, but of opposite sign, so it is possible that it switches back and forth as the reaction proceeds. As expected, most other significant residues are also in the vicinity of the QM part. Glu215, Val217, and Arg271 are located next to ligand residues His214, Asp216, and His270, respectively. Asn225 and Ser281 have H-bonds with the ACV substrate. Pro268 has H-bonds with ligand His270. Phe211 is close to the ligand oxygen molecule. Thr331 contacts to MM part of ACV (as shown in the Supporting Information) and Asn328 is a neighbor of Phe211 and His214. We also analyzed whether these geometrical effects are from electrostatic or VDW interactions. In above residues, only Asn252 in $XY =$ **56** and Gln225, Asn252, and Ser281 in $XY =$ **67** have strong geometrical effects because of their electrostatic interaction. Although there are many residues that have similar or even larger size of the electrostatic interaction changes, it is not enough to be strongest without the contribution from VDW. These Asn, Gln, and Ser residues have uncharged polar side chains. Residues with large electrostatic repulsion might not make enough VDW contact with the QM part. The geometrical effects of the other significant residues in Figure 10 are mainly caused by VDW interactions.

Figure 11a−d show S^{i}_{XY} for $XY =$ **23**, **67** (small statistical effects) and **45**, **56** (large statistical effects), respectively. Some residues located near the QM part appear in both the geometrical and statistical effects, but many others only contribute significantly to one of the effects. The statistical effects are mainly caused by VDW interactions, rather than electrostatic interactions because of the stronger distance dependence of the former. Bulky residues, like Phe211 and Pro268, in the vicinity of the QM part have large effects, while residues that contributed to the geometric effect by electrostatic interactions do not appear in the analysis of the statistical effect. Pairs $XY =$ **23** and **67** have relatively small geometric changes of the QM part. In $XY =$ **23**, C−H bond activation, the oxygen molecule moves slightly and the major movement is the hydrogen atom on the substrate ACV that moves closer to the oxygen. Among the significant residues in Figure 11a, Tyr189 and Ser281 both have H-bonds with QM part of ACV, while Phe211 is close to the oxygen molecule. Other significant residues, Phe41 and Pro268 are near the iron ligand His270, and Asn328 is next to the iron ligand His214. In the reaction corresponding to pair $XY =$ **67** (Figure 11b), the O−O distance of the peroxide formed after C−H activation increases in anticipation of O−O bond cleavage. Significant residues Phe221 and Asn252 are both close to the ligand water molecule that donates a proton during O−O bond cleavage, and Pro268 adjoin these two residues.

Both reactions with larger statistical effects (pairs $XY =$ **45** and **56**) also have large geometrical changes in the active site. In $XY =$ **45**, there is a rotation of the peroxide ligand away from the substrate and toward the ligand water that donate its proton. The significant residue Leu231 is located close to the oxygen molecule, while Phe41 is adjacent to the ligand water molecule and Leu231. Tyr189 and Ser281 both have H-bonds with the QM part of ACV, and Ile187 is next to these two residues. Arg271 is next to the ligand

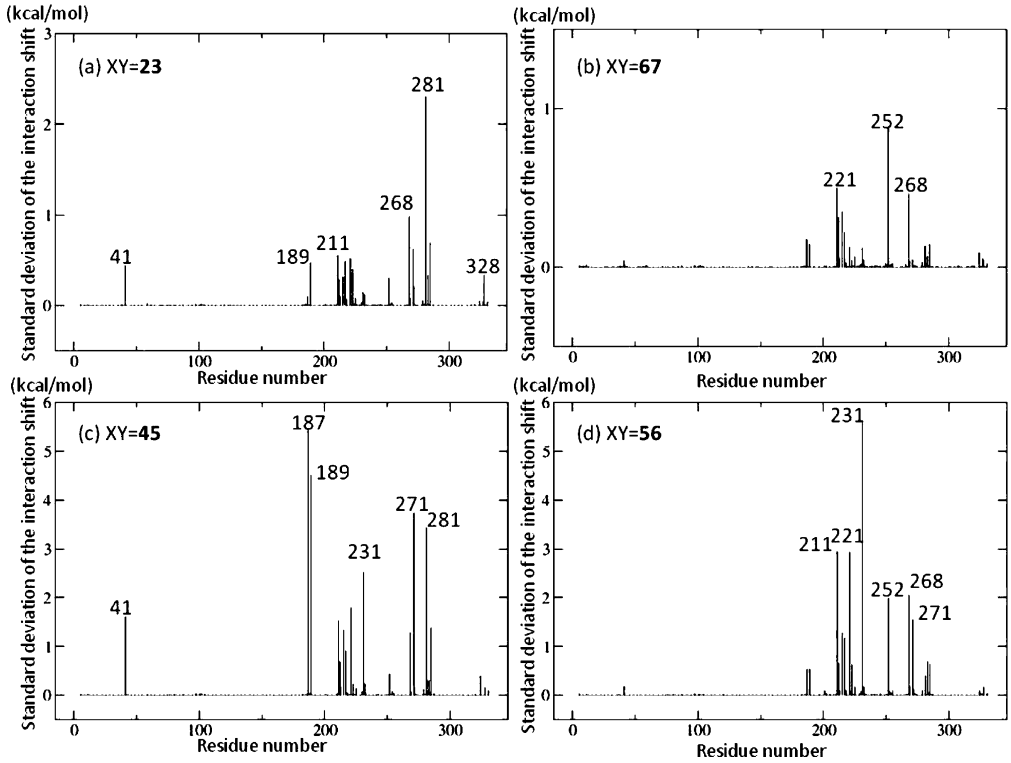


Figure 11. Standard deviation s^i_{XY} of nonbonded interaction change in MD simulation for individual residues, a representation of the statistical effect of the protein dynamics on the free energy profile. (a, b) Cases of weak statistical effect ($XY =$ **23** and **67**). (c, d) Cases of strong statistical effect ($XY =$ **45** and **56**). Note that there are different vertical scales for different figures.

His270 and adjoins the ligand His214 and another significant residue, Leu231. In $XY =$ **56**, an oxygen atom of the ligand OOH moves about 1 Å (another oxygen and hydrogen atoms also move ∼0.5 Å), and the ligand water rotates in the process of the electron transfer from iron to the oxygen molecule. Phe211, Leu231, and Arg271 are close to the ligand oxygen and Thr221 and Asn252 locate next of the ligand water. Pro268 is a neighbor of significant residue Asn252 as written above. Interactions between these residues and QM part significantly change in the reaction from state-**5** to state-**6**, and they contribute to the statistical effect of the free energy profile, see Figure 11d.

Comparing the values of the standard deviation in cases of weak statistical effect ($XY =$ **23** and **67**) within cases of strong statistical effect ($XY =$ **45** and **56**) in Figure 11, the size of the standard deviation for significant residues are much different (see vertical axes). On the other hand, significant values of $\Delta\Delta H^i_{XY}(X)$ are similar size in Figure 10a to 10d. These comparisons suggest that the geometrical effect is relatively stable for different reaction steps and the statistical effect can become large, because these values connect to the statistical and geometrical effects, respectively. When the dynamical contribution is large, the statistical effect dominates it in the reaction step. When the statistical effect is weak, sizes of the geometrical and statistical effects can be comparable.

4. Discussion

The goal in the present paper is to describe a method that can give a broad overview of dynamical contributions for complex multistep reactions. To achieve this, the most important approximations are the neglect of dynamical contributions on the QM region and a classical description of the interaction between QM and MM regions. With these two approximations, we avoid recalculating the QM wave function for each snapshot of the protein geometries.

The QM:[MM-FEP] approach belongs to a family of ONIOM approaches to model interactions between model and real system. ONIOM-ME describes these interactions classically and does not include polarization of the model system or the surrounding. In the semiclassical ONIOM-EE description, the model system is polarized by the charges of the surrounding, but polarization of the environment is not included. The geometry optimization procedure leads only to small changes in the protein structure, and does not really describe geometric polarization. The FEP approach includes a geometric polarization of the actual environment at a finite temperature but uses the classical representation of the electrostatic interactions.

Comparing the static and dynamical results of the potential diagram, the largest change is in the transition state barrier for O−O bond cleavage ($XY =$ **56**), see Figure 5. In this investigation, the dynamical contributions to the energetics have been separated into two types of effects, the geometric and the statistical effects. For the present reaction step, the statistical effect dominates the dynamical contribution of the protein on the free energy profile. The result suggests that the barrier at the QM:MM level includes an artificially high electrostatic repulsion, caused by lack of polarization in the frozen protein environment. The structure fluctuation of the protein environment screens this interaction and decreases the barrier height. We therefore suggest that when redox

center presents significant change of either geometry or electrostatic potential, the dynamical contributions to the free energy reaction diagram should be considered.

In FEP methods, all coordinates other than the selected reaction coordinate are integrated. Any type of the reaction coordinate can be selected, either a real geometry space coordinate or a virtual space coordinate. When choosing a real space coordinate, various computational developments have been reported using different integration methods for leaving coordinates.[47,48] In the most cases, one or a few reaction coordinates like bond, angle or dihedral torsion have been chosen. The approach is appropriate for discussing the realistic dynamics of the system. In the present investigation, we have employed the alchemical FEP method that uses a nonphysical coordinate. The reaction coordinate describes the appearance of atoms in the final state and disappearance of atoms in the initial state. One advantage is that this method can be used for any change of a system, not only chemical reactions. Here, we use the method out of convenience because it is possible to describe the reaction path between two intermediates without a detailed mapping of the potential energy surface. However, in cases where the appearing and disappearing atoms leads to significant changes in the Hamiltonian, the FEP calculation is sometimes hard to converge, and the alchemical FEP method has therefore been preferably applied to small fragments. The alchemical FEP method uses direct linear interpolation of the interaction potential energy. The potential interpolation for the free energy calculation is a traditional approach.[49] There is another approach to build a virtual reaction coordinate in the linear interpolation of the other physical values like atomic coordinates and charges.[50] This alternative approach still avoids the cost of determining the intermediate state and includes less approximation of the interaction energy. At same time, when the rotations of molecules are involved in the reaction, the direction of linear interpolation of the atomic coordinates must be properly chosen.

Our computed value of the dynamical contribution is not a well-determined component of the free energy, because it includes effects of the solvent, temperature difference (compared with 0 K model) and dynamics of the environment. The term is not for comparing with the nature or experiments, but for connecting QM:MM calculations to these. The geometrical effect includes both solvent and temperature effects. On the other hand, the statistical effect is a value connecting to the entropic energy. Computing the entropy in FEP criterion is challenging because the sampling of the correlation function of interactions is required instead of the sampling of interactions in the free energy calculation. That requires N-square order of the free energy sampling. The value of the statistical effect can be considered as an estimation of the entropic energy.

In the present study, we have selected a large fragment that includes the enzymatic reaction center as the alchemical part. The chemical reaction mainly occurs at the center of the fragment and only the metal ligands describe large geometric changes. We find that for reactions where the change in electronic structure is dominant, compared to changes in geometric structure, the alchemical FEP approach is well behaved. However, convergence is slow when strong VDW contacts appear during the FEP step. The error bars are largest for the release of water (step **89)** where the oxygen atom moves out about 1 Å from its original position in state **8** and occasionally makes strong VDW contacts with Leu231 and MM water molecules in the MD simulation. The same oxygen atom moves significantly (0.8−1.3 Å) also in other five reaction steps. But the difference is that in the final step, the water is released from the active site out into the environment. Water molecules are very flexibility in orientation inside the protein and such strong VDW contacts can be avoided. The perturbation criterion is broken at this moment and an artificial energy may be included in this part of FEP calculation. To avoid such problems, it might be necessary to check the change in the effective volume of the redox center when this method is applied.

Another problem appears when the statistical effect has a large contribution to the free energy profile. In that case the FEP calculation picks up only a small number of samples from the configurations that contribute the most. Therefore, such rare events get a very high weight in the FEP calculation. As an example $XY = \mathbf{56}$ has large fluctuation of ΔH_{XY} and the error bar of the computed dynamical contribution is also relatively large. For accurate calculations of reactions dominated by the statistical effect, longer sampling might be needed to include enough number of events, or it may be required to apply another approach.[51]

As we froze the redox center in the MM calculation, we neglect two free-energy terms caused by the dynamics of the redox center. One is that the average geometry of the redox center can change because of the protein dynamics. This effect can be included by applying free energy gradient methods.[52,53] The other term is the statistical effect of the redox center that influences the dynamics of neighboring residues and solvent. QM/MM sampling or an alternative method is required in order to include this term, which would result in a significant increase in computational cost. These terms depend on the computed system and model and are not clearly negligible. That is a challenging problem in the future.

5. Conclusion

The dynamical contribution of the protein environment to the reaction energy diagram has been included at the ONIOM QM:MM level, using an MM FEP method. We applied the method to eight reaction steps of the nonheme iron enzyme isopenicillin N synthase. Redox active metal enzymes require a relatively expensive QM treatment of the active site, and the method is therefore designed to avoid recalculating the QM density at any point of the dynamics simulation.

The dynamical contribution has been separated into geometric and statistical effects of the protein environment. The geometrical effect comes from a change in the average protein geometry and influences to the overall potential of the reaction diagram. The statistical effect is caused by the fluctuation of the interaction between MM and QM part during the molecular dynamics simulation. With respect to the IPNS enzymatic reaction mechanism, the inclusion of the dynamical contribution, mainly coming from the statisti-

cal effect, decreases the barrier for O−O bond cleavage by several kcal/mol. These results show that the dynamical fluctuations of the protein environment can be a factor when modeling enzymatic reactions.

Acknowledgment. One of the authors (M.L.) acknowledges a Fukui Institute for Fundamental Chemistry Fellowship. The work was in part supported by a CREST (Core Research for Evolutional Science and Technology) grant in the Area of High Performance Computing for Multiscale and Multiphysics Phenomena from the Japan Science and Technology Agency (JST). The use of computational resources at the Research Center of Computer Science (RCCS) at the Institute for Molecular Science (IMS) is acknowledged.

Supporting Information Available: Details of the modified RESP charge fitting procedure. Analysis of geometry changes during MD simulations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Que, L., Jr.; Tolman, W. B. *Nature* **2008**, *455*, 333–340.

(2) Siegbahn, P. E. M. *J. Biol. Inorg. Chem.* **2006**, *11*, 695–701.

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

(4) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(5) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959–1967.

(6) Matsubara, T.; Maseras, F.; Koga, N.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 2573–2580.

(7) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(8) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461−462*, 1–21.

(9) Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A., Jr.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815–826.

(10) Morokuma, K. *Proc. Jpn. Acad. Sci., Ser. B* **2009**, *85*, 167–182.

(11) Vreven, T.; Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701–704.

(12) Prabhakar, R.; Vreven, T.; Frisch, M. J.; Morokuma, K.; Musaev, D. G. *J. Phys. Chem. B* **2006**, *110*, 13608–13613.

(13) Lundberg, M.; Kawatsu, T.; Vreven, T.; Frisch, M. J.; Morokuma, K. *J. Chem. Theory. Comput.* **2009**, *5*, 220–234.

(14) Kwiecien, R. A.; Khavrutskii, I. V.; Musaev, D. G.; Morokuma, K.; Banerjee, R.; Paneth, P. *J. Am. Chem. Soc.* **2006**, *128*, 1287–1292.

(15) Li, X.; Chung, L. W.; Paneth, P.; Morokuma, K. *J. Am. Chem. Soc.* **2009**, *131*, 5115–5125.

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

(17) Yang, W.; Bitetti-Putzer, R.; Karplus, M. *J. Chem. Phys.* **2004**, *120*, 9450–9453.

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

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

(20) Zhang, X.; Bruice, C. T. *J. Am. Chem. Soc.* **2007**, *129*, 1001–1007.

(21) Kamiya, M.; Saito, S.; Ohmine, I. *J. Phys. Chem. B* **2007**, *111*, 2948–2956.

(22) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.

(23) Gao, J.; Kuczera, K.; Bruce, T.; Karplus, M. *Science* **1989**, *244*, 1069–1072.

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

(25) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2003**, *125*, 12035–12048.

(26) Rod, T. H.; Ryde, U. *Phys. Rev. Lett.* **2005**, *94*, 138302.

(27) Andersson, I.; Terwisscha van Scheltinga, A. C.; Valegård, K. *Cell. Mol. Life Sci.* **2001**, *58*, 1897–1906.

(28) Lundberg, M.; Siegbahn, P. E. M.; Morokuma, K. *Biochemistry* **2008**, *47*, 1031–1042.

(29) Lundberg, M.; Morokuma, K. *J. Phys. Chem. B* **2007**, *111*, 9380–9389.

(30) Pearlman, D. A.; Kollman, P. A. *J. Chem. Phys.* **1991**, *94*, 4532–4545.

(31) Pearlman, D. A. *J. Phys. Chem.* **1994**, *98*, 1487–1493.

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

(33) http://www.ks.uiuc.edu/Research/vmd/ (accessed Nov 13, 2010).

(34) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(35) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(36) 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*, development version;Gaussian, Inc.: Wallingford CT, 2008.

(37) Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani,

F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S. Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.

(38) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(39) Namd scalable molecular dynamics. http://www.ks.uiuc.edu/Research/namd/ (accessed Nov 13, 2010).

(40) Efron, B. *Annu. Stat.* **1979**, *7*, 1–26.

(41) Konishi, S. *Ann. Stat.* **1991**, *19*, 2209–2225.

(42) Ooura, T. Ooura's mathematical software packages. http://www.kurims.kyoto-u.ac.jp/~ooura/ (accessed Nov 13, 2010).

(43) Baldwin, J. E.; Abraham, E. *Nat. Prod. Rep.* **1988**, *5*, 129–145.

(44) Kriauciunas, A.; Frolik, C. A.; Hassell, T. C.; Skatrud, P. L.; Johnson, M. G.; Holbrok, N. I.; Chen, V. J. *J. Biol. Chem.* **1991**, *266*, 11779–11788.

(45) Johnson, B. G.; Gonzales, C. A.; Gill, P. M. W.; Pople, J. A. *Chem. Phys. Lett.* **1994**, *221*, 100–108.

(46) Patchkovskii, S.; Ziegler, T. *J. Chem. Phys.* **2002**, *116*, 7806–7813.

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

(48) Isralewitz, B.; Izrailev, S.; Schulten, K. *Biophys. J.* **1997**, *73*, 2972–2979.

(49) Mruzik, M. R. *Chem. Phys. Lett.* **1977**, *48*, 171–175.

(50) Zeng, X.; Hu, H.; Hu, X.; Yang, W. *J. Chem. Phys.* **2009**, *130*, 164111(1−8).

(51) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(52) Okuyama-Yoshida, N.; Nagaoka, M.; Yamabe, T. *Int. J. Quantum Chem.* **1998**, *70*, 95–103.

(53) Hu, H.; Lu, Z.; Parks, J. M.; Burger, S. K.; Yang, W. *J. Chem. Phys.* **2008**, *128*, 034105(1−18).

CT1005592

JCTC Journal of Chemical Theory and Computation

Quantitative Assessment of Force Fields on Both Low-Energy Conformational Basins and Transition-State Regions of the ($\phi-\psi$) Space

Zhiwei Liu,[†] Bernd Ensing,[‡] and Preston B. Moore*,[†]

West Center for Computational Chemistry and Drug Design, Department of Chemistry & Biochemistry, University of the Sciences in Philadelphia, 600 South 43rd Street, Philadelphia, Pennsylvania 19104, United States and Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Received July 16, 2010

Abstract: The free energy surfaces (FESs) of alanine dipeptide are studied to illustrate a new strategy to assess the performance of classical molecular mechanics force field on the full range of the ($\phi-\psi$) conformational space. The FES is obtained from metadynamics simulations with five commonly used force fields and from ab initio density functional theory calculations in both gas phase and aqueous solution. The FESs obtained at the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) level of theory are validated by comparison with previously reported MP2 and LMP2 results as well as with experimentally obtained probability distribution between the C_5-β (or β-PPII) and α_R states. A quantitative assessment is made for each force field in three conformational basins, LeRI (C_5-β-$C7_{eq}$), LeRII (β_2-α_R), and LeRIII(α_L-$C7_{ax}$-α_D) as well as three transition-state regions linking the above conformational basins. The performance of each force field is evaluated in terms of the average free energy of each region in comparison with that of the ab initio results. We quantify how well a force field FES matches the ab initio FES through the calculation of the standard deviation of a free energy difference map between the two FESs. The results indicate that the performance varies largely from region to region or from force field to force field. Although not one force field is able to outperform all others in all conformational areas, the OPLSAA/L force field gives the best performance overall, followed by OPLSAA and AMBER03. For the three top performers, the average free energies differ from the corresponding ab initio values from within the error range (<0.4 kcal/mol) to ~1.5 kcal/mol for the low-energy regions and up to ~2.0 kcal/mol for the transition-state regions. The strategy presented and the results obtained here should be useful for improving the parametrization of force fields targeting both accuracy in the energies of conformers and the transition-state barriers.

Introduction

Classical molecular dynamics (MD) simulations have become an indispensable tool to study the structure and dynamics of biological macromolecules, such as proteins, nuclear acids, and lipid assemblies.[1−7] Assuming the validity of the underlying classical approximation and sufficient sampling of the phase space, the accuracy of a MD calculation primarily depends on the quality of the molecular mechanics (MM) force field employed. Relentless effort to develop and parametrize MM force fields during the past three decades has led to the development of standardized force field libraries with brand names, such as the AMBER,[8] CHARMM,[9] GROMOS,[10] and OPLS models.[11] Considering the somewhat simple standard functional form of additive pair-potentials, (which include many-body effects, such as

* Corresponding author. E-mail: p.moore@usp.edu. Telephone: 215-596-7537.

[†] University of the Sciences in Philadelphia.

[‡] University of Amsterdam.

10.1021/ct100395n
Published on Web 12/29/2010

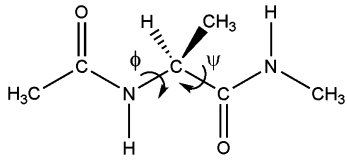


Figure 1. Structure of Ace-Ala-Nme (alanine dipeptide, AD).

polarization only in an effective manner), these force fields have been shown to reproduce experimentally determined equilibrium structures of biological macromolecules surprisingly well. However, with the constantly advancing computer power, the emphasis of MD simulations has gradually moved from structural prediction to dynamical and kinetic processes, such as protein folding.[12−19] The performance of the standard force fields in these areas of study has not yet been evaluated thoroughly. This is partially because the accuracy of a thermodynamic and/or kinetic process depends not only on a correct force field description of the equilibrium structure but also that of the conformational propensity between different conformers and transition states.

To investigate the conformational dynamics and kinetics of proteins, the alanine dipeptide (Ace-Ala-Nme, AD, Figure 1) molecule has been a standard model system for theoretical[16,20−29] and experimental studies for the past 20 years.[30−36] Experimentally, two-dimensional (2D) infrared (IR),[37] vibrational,[38] Raman,[39] vibrational circular dichroism,[40] and NMR spectroscopy[40] have been used to study the conformational preference of AD and other alanine peptides in aqueous solution. Theoretically, high-level ab initio methods, such as the second-order Møller−Plesset perturbation theory (MP2), have been used to calculate the potential energies of AD across the full range of the ($\phi-\psi$) conformational space, either in gas phase or using the implicit continuum solvent models.[25,41] The energy profile as a function of the backbone dihedral angles ϕ and ψ, known as the Ramachandran plot, has been used to parametrize and examine MM force fields[9,24,41−43] and in the development and evaluation of new modeling techniques.[44−46] Various enhanced sampling techniques, such as umbrella sampling,[47,48] adiabatic free energy dynamics (AFED)[49] method, and metadynamics,[50−52] have been developed and applied on mapping the ($\phi-\psi$) free energy surface (FES) of AD in both gas phase and aqueous solution. Most of these studies, with a focus on methodology development, have either compared FESs of AD among different force fields[52] or evaluated the force field FESs by high-level ab initio calculations in terms of positions (ϕ, ψ values) of energy minima, such as $C7_{eq}$, C_5, α_R, β, etc. and their relative stabilities. Both types of evaluation have limited abilities to give a full spectrum of assessment of the force field FESs. This is especially true for free energy in aqueous solution phase, even though experimental[27,37] and high-level ab initio data have become available.[25] Critically, no emphasis has been made on assessing the accuracy of transition states, which are essential in describing kinetics.

With respect to the dynamics, MD simulations of AD (and other small alanine peptides) up to 20 ns have been carried out using various MM force fields and semiempirical [such as self-consistent charge density functional tight binding (SCC-DFTB), PM3, and AM1] methods.[24,27,53,54] The performance of the force field has been evaluated by comparing the conformational propensity obtained from the force field and semiempirical MD simulations. However, with 20 ns simulations, only a part of the negative ϕ side of the ($\phi-\psi$) conformational space has been sufficiently sampled. Therefore, assessment of the force fields is made mostly on whether or not a proper ratio of population is achieved between the β and α_R conformers, not the full ($\phi-\psi$) space.

In the present work, we have applied the metadynamics method,[50,55] in combination with five different force fields, to explore the 2D ($\phi-\psi$) FESs of AD in both gas and aqueous solution phases. Further, we have mapped out the potential energy profiles of AD using ab initio method, specifically the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) combination, on a 50 × 50 grid for both gas and implicit solvent models. The quality of the B3LYP energy surfaces is carefully examined by comparing them with both experimental and high-level ab initio results, such as those using MP2 and LMP2 with fairly large basis sets.[25,56] The free energy corrections to the potential energy profiles are obtained from frequency calculations at each grid point. The assessment of the force field FESs was not done by simple comparison of geometry and energy of each individual minimum- or transition-state structure against the ab initio result. Instead, the 2D ($\phi-\psi$) map was divided into three low-energy conformational basins and three transition-states regions. Quantitative assessment of each force field was carried out by two comparisons between force field and ab initio results: (1) the average free energy of each region; (2) how well a force field FES would match the ab initio FES in each region by calculating the standard deviation of the free energy difference between the two FESs.

Methods

Ab Initio Calculations. Ab initio energy surfaces of AD were computed on a 50 × 50 grid in the ($\phi-\psi$) conformational space in which each dihedral angle had a range of −180° to 180° with a 7.2° interval. At each grid point, geometry parameters, except the constrained ($\phi-\psi$) dihedral angles, were fully optimized at the B3LYP/6-31G(d,p) level of theory.[57−59] For AD in water, the optimization was performed in a polarized reaction field. The dielectric constant of water and the polarized continuum model (PCM)[60−62] implemented in Gaussian 03[63] were used. All energy minima and transition-states found on the constructed energy profiles were further fully optimized by removing the $\phi-\psi$ constraints. The potential energy of AD at each grid point, minimum or transition state, was then refined by single point calculation at the B3LYP/6-311+G(2d,p) level of theory. Frequency calculation at the B3LYP/6-31G(d,p) level of theory was carried out on each partially (grid point) or fully optimized structure to characterize the minimum (no imaginary frequency) or the transition state (one imaginary frequency) and to obtain the free energy corrections, including zero point energies (ZPE) (scaling factor of 0.9804)[64] and thermal energy (enthalpy and entropy) corrections at 298.15 K and 1 atm.

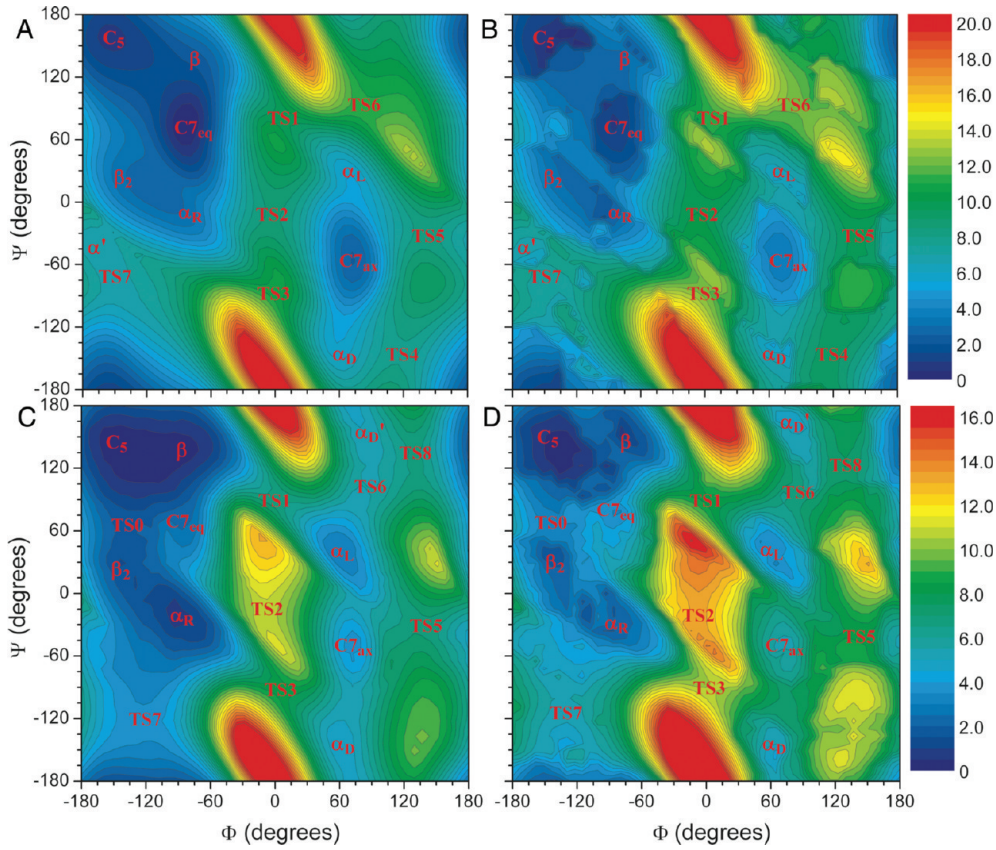


Figure 2. B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) energy contour maps of AD. (A) Potential energy (relative to $C7_{eq}$) in gas phase. (B) Free energy (relative to $C7_{eq}$) in gas phase. (C) Potential energy (relative to C_5) in aqueous solution phase. (D) Free energy (relative to C_5) in aqueous solution phase.

Metadynamics Simulations. All metadynamics simulations were carried out using the CM3D program.[65] Five different force fields, CHARMM27,[9,41] AMBER94,[8] AMBER03,[42] OPLSAA,[11] and OPLSAA/L (modified torsion parameters),[43] were used to describe AD and the water molecules, with the exception of using SPC/f for the water model.[66,67] All of these force fields use pair potentials that have been parametrized using a combination of quantum mechanical and experimental data to reproduce the minimum structures of macromolecular molecules, such as proteins, RNA, and DNA. These force fields also have similar functional forms (bonded and nonbonded terms). While there are no gross differences, they do differ in the details of parametrization. For example, AMBER03 is reparameterized using new point charges and backbone torsion parameters from AMBER94 to improve the performance of condensed-phase simulations. Similarly, OPLSAA/L is an improved version of OPLSAA by refitting the backbone torsion parameters. As a result, specific molecular systems will show differences, such as in simulations of membranes[68] or DNA.[69] It should be noted that the cross terms (CMAP) in the CHARMM27 force field were not used to evaluate the force fields within the boundary of the standard functional forms. For the simulation of aqueous solution, an AD molecule was placed in a periodic cubic box ($L = 18.8$ Å) with 216 water molecules. The electrostatic interactions were calculated using Ewald summation,[1,70,71] and the real space cut-off was half of the cell dimension (9.4 Å). Prior to the metadynamics runs, NPT simulation at 1 atm and 298 K for at least 100 ps was carried out to equilibrate the cell volume. The metadynamics simulations were carried out using the NVT ensemble and a time step of 0.5 fs.

We used two sets of parameters related to the Gaussian potentials (or "hills"):[50,51] (1) $w = 0.2$ rad, $h = 0.02$ kcal/mol, and $\Delta = 0.02$ rad and (2) $w = 0.1$ rad, $h = 0.02$ kcal/mol, and $\Delta = 0.075$ rad. Here, Δ is the minimum distance that the systems must move in $(\phi-\psi)$ space before the next Gaussian potential is added. Hills setting (2) thus employs smaller potentials that are also further away from each other than hills setting (1). It should increase the accuracy of the FES, however, it takes longer simulation time to sample the entire conformational space. For AD in gas phase, both settings were employed. For hills setting (1), a 5 ns productive run was carried out to achieve sufficient sampling of the whole $(\phi-\psi)$ conformational space. For hills setting (2), 20 ns are needed to achieve reasonably sufficient sampling for most of the force fields. The reconstruction of the $(\phi-\psi)$ FES was done on a 50 × 50 grid, in which the grid interval of 7.2° was comparable with the width of the hills in both settings. The same grid resolution was used throughout the paper so that the energy difference map between any two FESs, which is essential for our quantitative analysis, can be easily calculated.

Results and Discussions

Highlights. Results obtained from this study fall into three subsections: (1) ab initio FESs of AD in gas and aqueous

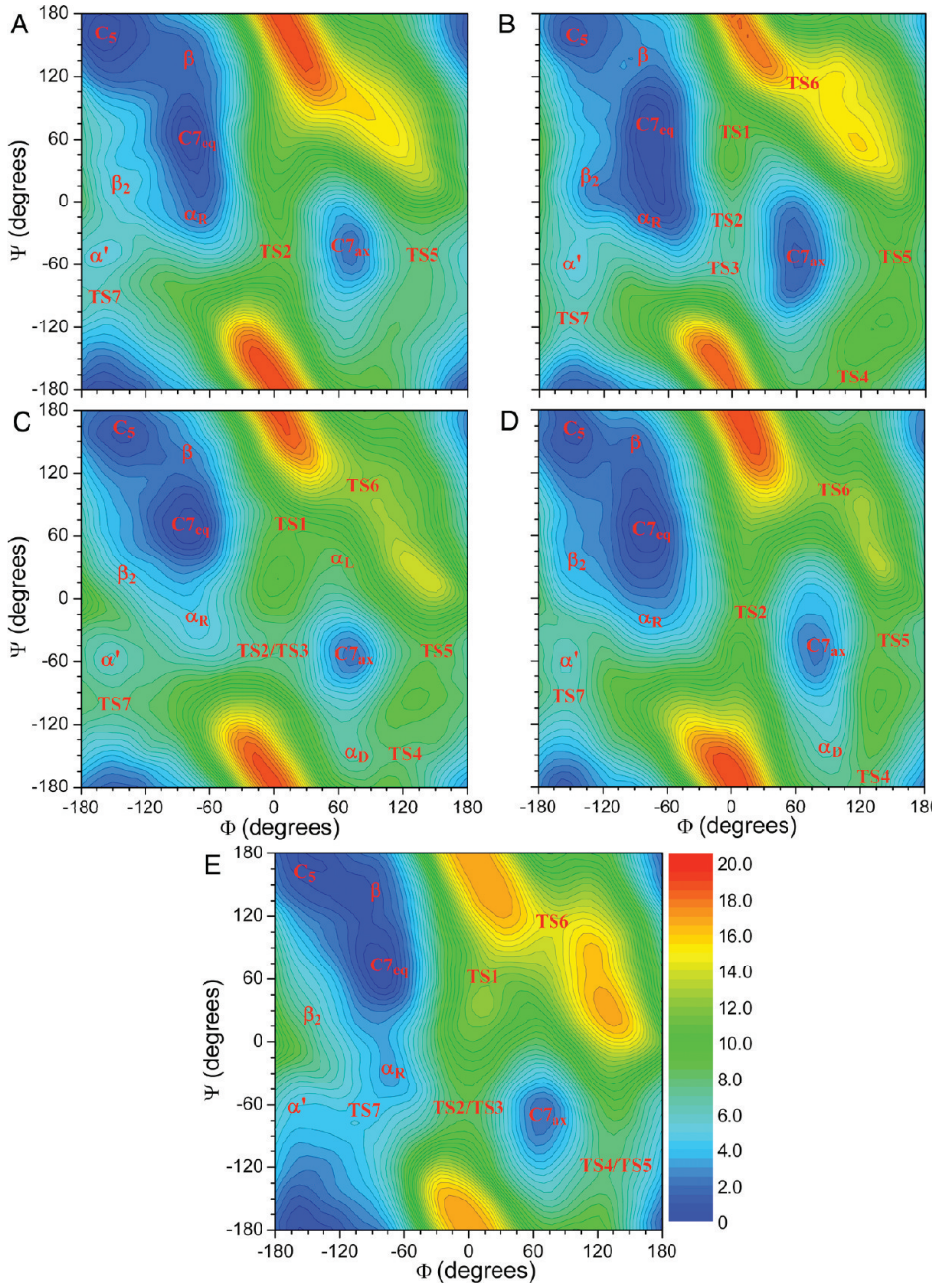


Figure 3. Free energy contour maps of AD in gas phase: (A) AMBER03; (B) AMBER94; (C) OPLSAA; (D) OPLSAA/L; and (E) CHARMM27.

phases; (2) FESs of AD obtained by metadynamics simulations using the five force fields; (3) quantitative assessment of the force fields according to the ab initio FESs. Part (3), though comes at the end, is the most important part of this study. Therefore, to give the readers a synopsis of what this study concludes, here we are providing the highlights of our results, which mainly focus on part (3), followed by more detailed discussion of each subsection.

Our study shows that the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) method (with PCM for aqueous phase) is able to generate energy contour maps of AD in gas (Figure 2A) and aqueous (Figure 2C) phases that are as accurate as the previous MP2 approaches.[25,41] The computationally cost-efficient hybrid density functional theory (DFT) method also allowed us to perform frequency calculation and obtain free energy corrections at each grid point so that the ab initio FESs (Figure 2B and D) can be constructed. Second, metadynamics simulations in combination with several commonly used force fields, namely AMBER94, AMBER03, CHARMM27, OPLSAA, and OPLSAA/L, are used to generate FESs of AD in gas (Figure 3) and aqueous (Figure 4) phases (explicit solvents). The estimated errors of the FESs are 0.2−0.4 kcal/mol for the ab initio FESs and 0.3−0.4 kcal/mol or slightly larger (up to 0.7 kcal/mol) in some cases due to minor insufficient sampling for the force field FESs.

The ab initio FESs are used as standards to assess the accuracy of the MM force field FESs in the low-energy conformational basins and critically for transition states linking the minima. Specifically, the ($\phi-\psi$) conformational space has been divided into six regions (Figure 5). The

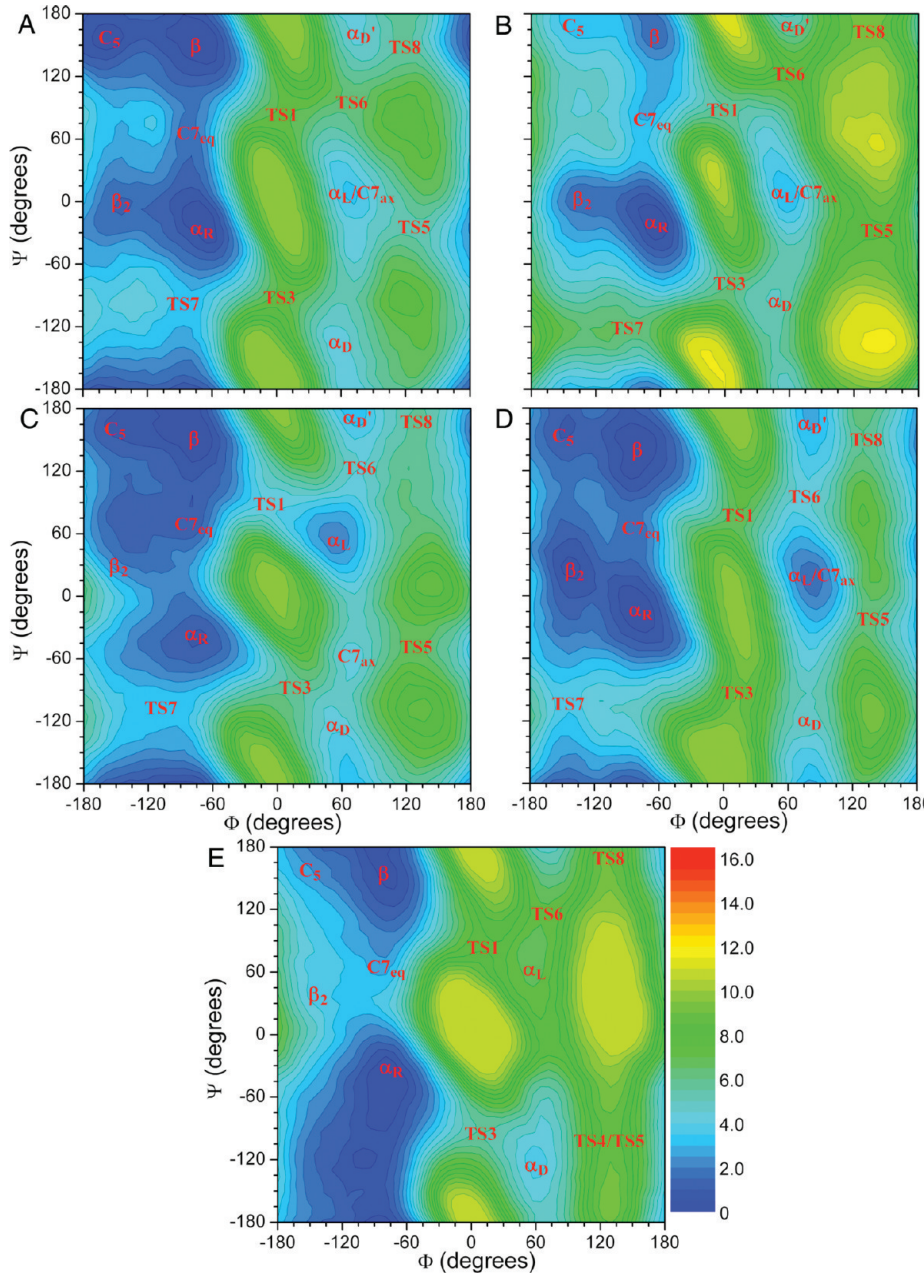


Figure 4. Free energy contour maps of AD in aqueous solution. (A) AMBER03; (B) AMBER94; (C) OPLSAA; (D) OPLSAA/L; and (E) CHARMM27.

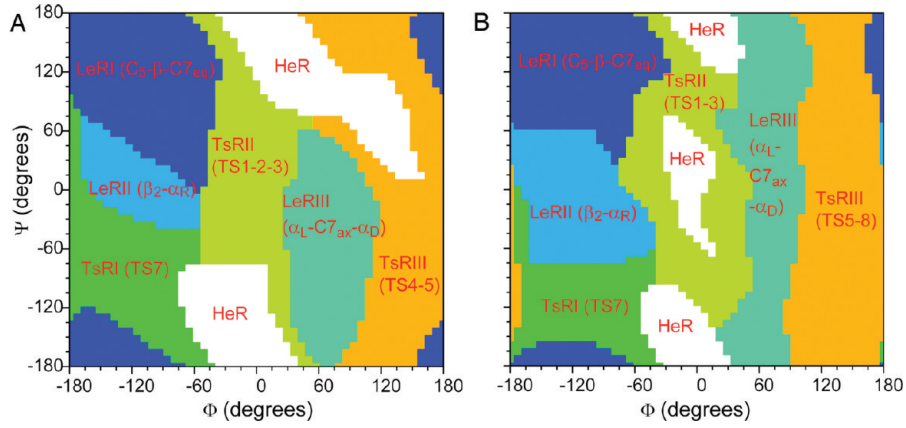


Figure 5. Definitions of low-energy (LeR) and transition-state (TsR) regions: (A) gas and (B) aqueous solution phases.

partitions are based on common practice of protein backbone conformational subsets as well as the ab initio FESs. There are three low-energy conformational basins (denoted as LeR), in which LeRI encloses minima C_5 (corresponding to β in some literatures), β (corresponding to PPII in some literatures), and $C7_{eq}$. The LeRII basin includes minima β_2 and α_R, and both the LeRI and LeRII regions are on the negative ϕ side. LeRIII is on the positive ϕ side and covers minima α_L, $C7_{ax}$, and α_D. There are also three transition-state regions, in which TsRI is the low-lying transition region between LeRII and LeRI. Both TsRII and TsRIII link the negative (LeRI and LeRII) and positive (LeRIII) ϕ sides through rotation of the ϕ dihedral angle clockwise (through 0°) or counterclockwise (through 180°), respectively. In this study, we have measured two quantities (Tables 3 and 4) per region for any FES. The first one is the mean free energy (average free energy of all grid points included in one basin), and the other is the standard deviation of the free energy difference map (FEDM) (obtained by one to one subtraction at each grid point between two FESs) of the designated region between the force field and the ab initio FESs. Comparison of the mean free energy of each basin (relative to the overall mean of all six regions on a FES) between ab initio and force field roughly indicates the accuracy of free energies of a particular group of conformers or transition states. The standard deviation of the FEDM measures how well the two FESs match each other and therefore provides a good indication of whether or not the force field predicts the positions of minima or transition states accurately. The following paragraphs give a brief summary of the assessment for each force field employed. It is worth mentioning that the average free energy of LeRII (β_2-α_R) is 0.9 kcal/mol higher than that of LeRI (C_5-β-$C7_{eq}$) in aqueous solution based on our ab initio FES. This free energy gap is in agreement with a probability partition of 80−20 between the C_5-β and the α_R states, as reported in recent experimental/computational investigations.[27,38]

For AMBER03, for the gas phase, the mean free energies of most regions agree well with the ab initio values. However, the LeRII (β_2-α_R) stability is slightly underestimated (0.9−1.2 kcal/mol higher). The energy surfaces match better with the ab initio FES on the negative ϕ side (deviations ∼1 kcal/mol) than on the positive ϕ side (deviations ∼1.5 kcal/mol). For aqueous solution phase, all three low-energy regions have higher, while all three transition-state regions have lower, mean free energies than ab initio. This means that the energy barriers for transition between conformational basins are 1.2−2.4 kcal/mol underestimated. In terms of the standard deviations, the AMBER03 aqueous FES matches well in all regions with the ab initio FES except for TSRI and TsRII, which have slightly larger deviations (1.5−1.7 kcal/mol). In addition, the aqueous phase FES matches with ab initio slightly better than the gas-phase FES, which reflects the parametrization emphasis of AMBER03 for condensed phase.

For AMBER94, for both gas and solution phases, the LeRI (C_5-β-$C7_{eq}$) stability is underestimated. The effect of this underestimation is severe for the aqueous solution phase FES since it has caused a reversed order of stability between LeRI (C_5-β-$C7_{eq}$) and LeRII (β_2-α_R). In addition, the mean free energy is lower for TSRII and higher for TSRIII than ab initio for both gas and aqueous phases, but the differences in energy are bigger in aqueous solution. The energy barriers decrease up to 5.5 kcal/mol for TsRII and increase 1.3 kcal/mol for TsRIII, which will affect any kinetic model built based on the AMBER94 force field.

CHARMM27 underestimates the stability of LeRII (β_2-α_R) in gas phase but not in aqueous solution and vice versa for LeRIII (α_L-$C7_{ax}$-α_D). It also gives a higher mean free energy for LeRI (C_5-β-$C7_{eq}$) and a lower mean for TsRII in aqueous solution, thus lowering the barrier 1.8−2.8 kcal/mol. In addition, CHARMM27 also heavily overestimates the stability of TsRI, i.e., the lower left quadrant of the (ϕ−ψ) conformational space in both gas and aqueous phases. The standard deviations in all regions are generally the first or second highest among all force fields. These larger deviations from ab initio are probably the result of the CHARMM parametrization procedure, which includes more empirical adjustments to fit with experimental crystallographic data.[9]

The performances of OPLSAA and OPLSAA/L are generally similar, and both have lower standard deviations (0.7−1.5 kcal/mol) in all six regions comparing with other force fields. For both gas and aqueous phases, the OPLSAA mean free energies of LeRII (β_2-α_R) are higher than the respective ab initio values. OPLSAA/L improves the average free energies of LeRII significantly. However, OPLSAA gives a much better relative stability of LeRI (C_5-β-$C7_{eq}$) versus LeRII (β_2-α_R) due to similar underestimation of stabilities of both regions for the aqueous phase. Conversely, the improvement of OPLSAA/L in LeRII (β_2-α_R), in combination with the underestimation of stability of LeRI, leads to too small of an energy difference between LeRI and LeRII for aqueous phase. Both force fields also have lower mean free energies for TsRII, which leads to a 3.0 kcal/mol decrease in barrier for OPLSAA and a 2.4 kcal/mol for OPLSAA/L.

In summary, in our opinion, OPLSAA/L gives the best performance overall, followed by OPLSAA and AMBER03. As recently pointed out by Feig,[72] force fields parametrized based on AD can accurately reflect amino acid backbone torsional preferences when used in MD simulations of proteins. Therefore, the quantitative assessment and strategy presented here can be used for improving force field parametrization targeting not only the accuracy of energies of conformers but also transition-state barriers. It should be noted that the torsional profiles of AD obtained should not be directly used to represent the torsional preference of amino acids in protein structures.[72]

Free Energy Surfaces of AD by Ab Initio Calculations. Previously, various ab initio calculations at different levels of theory have been used to study AD in gas phase, water, and/or other medium, in order to gain insights into the conformational preferences of protein backbones and/or in assisting parametrization of MM force fields.[41,56,73,74] Particularly, Wang et al. have used the MP2/cc-pVTZ//MP2/6-31G(d,p) level of theory, in combination with the PCM model, for solvation effects to obtain fully relaxed (ϕ−ψ) potential energy surfaces of AD.[25] In addition to fully relaxed

Table 1. Optimized Geometries (ϕ and ψ in °) and Relative and Free Energies (ΔE and ΔG in kcal/mol) of Stationary Points of AD in a Gas Phase, Obtained at the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) (this work) and MP2/cc-pVTZ//MP2/6-31G(d,p)[25] Levels of Theory

	B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p)				MP2/cc-pVTZ//MP2/6-31G(d,p)			
	ϕ	ψ	ΔE	ΔG	ϕ	ψ	ΔE[c]	ΔG
$C7_{eq}$	−83.1	72.6	0.00	0.00	−82.0	80.6	0.0	0.00
C_5	−158.4	164.6	0.92	0.06	−159.7	159.3	1.47 (0.91−1.11)	0.67
α_R	−80.0	−20.0	3.37[a]	2.28	−80.0	−20.0	3.27	3.74
α_L	68.4	26.5	5.48	5.19	63.2	35.4	4.52 (4.36−5.19)	4.99
$C7_{ax}$	73.6	−57.7	2.48	2.63	75.8	−62.8	2.50 (2.06−2.48)	2.32
β_2	−125.7	21.6	2.72	1.87	−141.6	23.8	3.25 (2.51−2.84)	2.63
α'	−169.9	−39.2	6.59	5.88	−166.1	−36.7	6.07 (5.49)	6.15
α_D	59.8	−136.2	5.53[b]	4.55	53.0	−133.4	4.75	5.24
TS1	5.6	81.4	9.72	10.31	4.4	84.3	8.94	
TS2	−1.4	−8.9	8.64	8.94	−0.2	−26.2	9.37	
TS3	2.8	−77.3	10.68[b]	10.39	4.4	−85.5	10.20	
TS4	112.8	−146.7	7.93	8.24	115.4	−151.9	8.44	
TS5	135.9	−26.2	8.17	8.68	135.0	−25.3	8.12	
TS6	79.0	86.4	11.23	11.04	75.2	90.7	11.20	
TS7	−149.8	−87.3	6.80	6.27				

[a] Partially optimized by constraining (ϕ,ψ) at (−80.0°, −20.0°). [b] Located by a series of partial optimization. [c] Values in parentheses are from ref 56 at levels ranging from LMP2/cc-pVTZ(-f)//MP2/6-31G* level to LMP2/cc-VQZ(-g)//MP2/6-311++G**.

energy surfaces in the gas phase, ether and water, they have also optimized and characterized all energy minima and transition states as well as calculated free energy corrections of energy minima. Mackerell et al. have constructed the potential energy surface of AD in gas phase at a similar level of theory.[41] In addition, they and others have also optimized a few energy minima in the gas phase using levels up to MP2/6-311++G(d,p). In this paper, we have utilized a more time-efficient hybrid DFT method, specifically, the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) method, in combination with the PCM model for solvation effects. Our energy contour maps were obtained in a similar fashion with a finer grid. Furthermore, we have also estimated the free energy corrections to the energy contour maps. Finally, we optimized and characterized all energy minima and transition states. The comparison with the previous results will not only give us an estimate on the accuracy of the more time-efficient DFT methods but also test the robustness of the approach.

Figure 2A and B shows the potential energy surface and the FES of AD in the gas phase. The general features of the potential energy map are very similar to an energy map calculated at the MP2/cc-pVTZ//MP2/6-31G(d,p) level of theory.[25] The positions and numbers of energy minima and transition states on the two energy maps coincide with each other. Particularly, both the MP2 and DFT energy map indicate a transition state (TS2) at around (0°, 0°), which is different from the HF map where only the energy maximum is observed between TS1 and TS3. This has indicated that the DFT method is able to describe the electrostatic interactions between the C═O and N−H groups in AD, when compared to the HF method. Full geometry optimization at the B3LYP/6-31G(d,p) level has located energy minima $C7_{eq}$, C_5, $C7_{ax}$, β_2, α_L, and α' as well as transition-states TS1, TS2, and TS4−TS7. Similar to all previous theoretical studies, minimum α_R cannot be located in gas phases, therefore optimization with ($\phi-\psi$) constrained at (−80°,−20°) is carried out. For α_D and TS3, no stationary points can be definitely located at the B3LYP/6-31G(d,p) level because the energy derivative with respect to the ψ dihedral angle does not converge to zero. However, we have located the best approximation of α_D and TS3 by a series of partial optimizations (detailed in Supporting Information) as (59.8°, −136.2°) and (2.8°, −77.3°), respectively. The energy gradients with respect to ψ are estimated at less than 0.008 kcal/mol deg.

As shown in Table 1, the optimized structures, potential and free energies of minima, and transition states agree fairly well with previous theoretical results. The relative potential energy order of all energy minima $C7_{eq}$, C_5, $C7_{ax}$, β_2, α_R, α_L, α_D, and α' is the same between the DFT and MP2 methods. Quantitatively, the differences in optimized dihedral angles ϕ and ψ between the two methods are mostly about 0−9° with two exceptions (β_2 and TS2) at around 15−17°. This is understandable since both minimum β_2 and transition state TS2 lie on the flat regions of the energy landscape. In fact, theoretical studies from different sources have shown significant discrepancy between optimizations using the MP2 method with different basis sets. Specifically, the β_2 conformer was optimized at (−141.6°, 23.8°) with MP2/6-31G(d,p),[25] (−125.7°, 21.6°) with B3LYP/6-31G(d,p), and (−90.7°, −7.8°) with MP2/6-311++G(d,p).[41] Interestingly, the optimized ϕ angle value determined by the DFT method is closer to that from the MP2/6-311++G(d,p) level with a much larger basis set than the MP2/6-31G(d,p) value with the same basis sets. The results indicate that the sizes of the basis sets (perhaps the addition of the diffusive functions) affect the optimized geometry more than the difference in methodology. Therefore, we are confident that the DFT method performance is at least as good as the MP2 method in terms of geometry optimization.

The differences in relative potential energies ($C7_{eq}$ as reference zero) of α_R, $C7_{ax}$, TS5, and TS6 are within 0.1 kcal/mol between the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) and MP2/cc-pVTZ//MP2/6-31G(d,p) methods. The relative energies of C_5, β_2, TS2, and TS4 from the DFT calculations are ∼0.5−0.7 kcal/mol less than that of the MP2 results. While the relative energies of α_L, α', α_D, TS1, and TS3 from the DFT calculations are ∼0.5−0.9 kcal/mol more

than that of the MP2 results. Generally, the DFT method predicts more stable extended structures (C_5, β_2) and less stable compact conformers (α_L, α_D) than the MP2 method. It has been recognized that the MP2/cc-pvTZ//MP2/6-31G** may artificially stabilize the compact conformers with respect to the extended ones due to intramolecular basis set superposition errors (BSSE).[56] The current energies calculated at the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) could be more accurate than that of the MP2/cc-pVTZ//MP2/6-31G(d,p) level. Furthermore, Table 1 has also listed the relative energies of C_5, α_L, β_2, $C7_{ax}$, and α' with respect to $C7_{eq}$ at the levels up to LMP2/cc-pVQZ(-g)//MP2/6-311++G(d,p).[56] Since the local MP2 (LMP2) method can circumvent the BSSE issue, the fact that the LMP2 energies agree better with the DFT method (on C_5, β_2, and in one case α_L) further confirms the above conclusion.

The free energies of all energy minima also agree fairly well, with a difference generally smaller than 0.7 kcal/mol, between the DFT and MP2 methods, with an exception of conformer α_R which has a free energies difference of 1.5 kcal/mol between the two methods. This difference, however, has no indication on the accuracy of the employed DFT method based on the fact that Wang et al. have estimated the α_R free energy corrections using the α_L conformer, while we have calculated them using the α_R conformer. Apart from the α_R conformer, the DFT method also tends to produce more stable extended structures (C_5, β_2, α', α_D) and less stable compact conformers (α_L, $C7_{ax}$) in terms of free energies relative to $C7_{eq}$, when compared to the MP2 method. The contributions to this trend come from either the potential energies (C_5, β_2,and α_L) or the free energy corrections (α', α_D, and $C7_{ax}$). The free energy of the C_5 conformer is 0.06 kcal/mol higher than the $C7_{eq}$ conformer at the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) level and 0.67 kcal/mol at the MP2/cc-PVTZ//MP2/6-31G(d,p) level. Both results agree well with the experimental observation that C_5 and $C7_{eq}$ were the dominant species at room temperature in the CCl_4 solution and at low temperature in Ar matrices.[75]

Although the discrepancies between the two approaches are not large enough to raise serious concerns, we do want to note two facts that favor the approach in the current paper. One is that the frequency calculations and the geometry optimizations in this paper are done at the same B3LYP/6-31G(d,p) level of theory, while in the previous case, HF/6-31G(d) was used to optimize the structures and calculate the frequencies, and the free energy corrections obtained were then applied on optimized structures at the MP2/6-31G(d,p) level of theory. The second is that the current calculations are performed with tight optimization criteria and an ultra fine grid for integration.

The FES of AD in gas phase, as shown in Figure 2B, is constructed on the basis of potential energy map plus free energy corrections, including zero point energies and thermal energy corrections based on frequency calculations. The accuracy of the free energy corrections is subject to the magnitudes of the remaining nonzero first derivatives (force) of the optimized structures at each grid point. Generally, there are two types of errors. One behaves more like noises since it comes from the uneven qualities of the optimized structures at each grid point, i.e., differences in remaining forces below the convergence criteria. Other sources of inaccuracy, such as vibrational−rotational coupling and hindered rotors, may also contribute to the noise of the FES. We have applied a linear smoothing function to even out noises. Specifically each grid point has only 50% contribution from the frequency calculation at this grid point, the other 50% comes evenly from its four neighboring grid points (ϕ −7.2°, ϕ +7.2°, ψ −7.2°, and ψ +7.2°). This technique works quite well based on the comparison of the FESs before and after the smoothing. The second type of error is more systematic since it comes from the remaining forces (nonharmonic character outside the minima or transition states) caused by the constraints of the two dihedral angles. We have also analyzed the behaviors and the magnitudes of the second type of systematic errors. Naturally, the free energy corrections are most accurate around the true stationary points, i.e., minima and transition states. The frequency calculations lead to underestimation or overestimation of the free energy corrections in the area that extends out from a minimum or a transition state, respectively. By analyzing the trends of free energy corrections along the paths from minimum to transition state, we can estimate that the maximum error would be around 0.6−0.7 kcal/mol, while the average error would be around 0.2−0.4 kcal/mol.

The overall features of the FES are similar to those of the potential energy surface as expected. Although the free energy corrections do not alter the positions of energy minima and transition states, they do slightly increase the slopes of the energy profiles, which results in more distinguished local energy minima, such as the separation of the β_2-α_R area from the C_5-β-$C7_{eq}$ area or α_L and α_D from the $C7_{ax}$ center. The average free energy correction to all the transition states is 0.7 kcal/mol higher than the average correction to all the minima, as also shown in Table 1.

The fully relaxed energy maps without and with free energy corrections of AD in water, obtained at the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) level with the PCM model, are shown in Figure 2C and D, respectively. The energy map in Figure 2C includes the potential energy of solute and polarized solute−solvent (PS-S) interaction energy, while the free energy corrections include zero point energies and thermal (enthalpy and entropy) corrections to the solute from the frequency calculation as well as the nonelectrostatic energy term, i.e., cavitation, dispersion and repulsion energies, calculated by the PCM model. The terms, energy maps without and with free energy corrections, are used in the following discussion with regard to Figure 2C and D, respectively.

The topological features of the energy map and changes from the gas-phase energy map to the aqueous map are almost identical to that obtained at the MP2/cc-pVTZ//MP2/6-31G(d,p) level with PCM model for solvation effects.[25] The changes include: (1) expansion of low-energy regions, more accessible and flatter energy surface (note the different contour color scale for gas (0−20 kcal/mol) and solution (0−16 kcal/mol) phases in Figure 2); (2) diminishing of the gas phase global minimum $C7_{eq}$ and emerging of the new minima α_R and β; (3) energy barrier (TS0) between C_5-β

Table 2. Optimized Geometries (ϕ and ψ in °) and Relative and Free Energies (ΔE and ΔG in kcal/mol) of Stationary Points of AD in Water, Obtained at the B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) (this work) and MP2/cc-pVTZ//MP2/6-31G(d,p)[25] Levels of Theory

	B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p)				MP2/cc-pVTZ//MP2/6-31G(d,p)			
	ϕ	ψ	ΔE	ΔG	ϕ	ψ	ΔE	ΔG^b
$C7_{eq}$	−85.4	73.4	2.06	3.09	−86.3	90.1	0.92	2.42
C_5	−151.6	147.6	0.00	0.00	−156.4	143.8	0.00	0.00
α_R	−78.1	−27.2	0.84	1.60	−70.5	−32.1	0.08	0.94
β	−75.1	143.3	0.26[a]	0.06	−64.0	142.1	0.17	0.39
α_L	61.3	40.9	2.84	3.30	59.4	41.4	1.27	1.67
$C7_{ax}$	73.4	−53.0	3.48	4.32	74.9	−54.3	2.69	4.00
β_2	−138.5	27.3	1.36	1.37	−145.6	27.2	1.27	1.57
α_D	60.1	−147.7	4.43	4.30	55.6	−144.9	3.59	3.57
$\alpha_{D'}$	72.8	164.8	4.79	4.01	60.5	−170.9	4.08	3.68
TS0	−129.8	62.6	1.74	2.63	−143.1	70.9	1.80	
TS1	0.3	91.6	6.41	7.88	0.1	91.4	5.83	
TS2	−11.0	−11.7	10.55	12.52	−6.6	−29.6	10.70	
TS3	7.3	−92.1	8.08	9.99	7.5	−92.2	7.50	
TS5	132.2	−28.1	6.58	7.78	130.4	−28.7	6.51	
TS6	81.3	104.2	5.28	6.45	76.4	104.5	6.30	
TS7	127.9	133.7	6.45	7.09				
TS8	−114.0	−115.6	3.44	4.99				

[a] Located by a series of partial optimization. [b] Free energies corrections were made at the HF/6-31G* level.

and α_R-β_2 regions; and (4) the shift of dominant region from $C7_{ax}$ to α_L. The optimized geometries and relative energies of all energy minima and transition states of AD in water are also shown in Table 2. The agreement on both the optimized (ϕ, ψ) angles and energies is again reasonable, indicating the robustness of both ab initio treatments. The topological feature changes from the gas phase to the aqueous map bring the aqueous map into closer agreement with the experimental Ramachandran plot derived from experimental protein structures.[76] The results indicate that the PCM approach gives good description of the bulk solvent polarization effects despite the lack of specific interactions between the solute and solvent molecules.

The discrepancies in optimized geometries between DFT and MP2, however, are slightly larger than that from the gas-phase calculations. This directly results from more extended and flatter low-energy regions in the aqueous energy surface. There are five stationary points ($C7_{eq}$, β, α_D', TS0, and TS2), instead of the two in the gas phase, whose optimized ϕ or ψ angles differ by more than 10° but no more than 25° between the DFT and MP2 results. For relative energies (C_5 as reference) without the corrections, the two methods agree very well (within 0.1 kcal/mol) on extended structures like β and β_2. However, for more compact structures, such as $C7_{eq}$, α_L, α_R, and $C7_{ax}$, the DFT method gives higher energies (relative to C_5), roughly 0.8−1.5 kcal/mol higher than that from the MP2 method. The discrepancies are systematic since relative stabilities (e.g., relative energies to $C7_{eq}$) among the more compact structures are similar (difference 0.3−0.4 kcal/mol) between the two methods. This is consistent with what we have observed in the gas-phase calculations, which states that BSSE errors at the MP2/cc-pVTZ//MP2/6-31G(d,p) level might have artificially stabilized more compact structures. For energy without corrections, both methods predict C_5 as the global minimum. The orders of energies for the minima are not exactly the same but very similar. Due to the stabilization on the extended structures by the DFT method, the β and β_2 conformers switch their places with their respective neighbors, the α_R and $C7_{eq}$ conformers, in the energy order (low to high) C_5-α_R-β-$C7_{eq}$-β_2-α_L-$C7_{ax}$-α_D-α_D' from the MP2 method. As a result, the energy order of C_5-β−α_R-β_2-$C7_{eq}$-α_L-$C7_{ax}$-α_D-α_D' is obtained at the level of B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p).

Similar to the MP2 study,[25] the free energy corrections have also changed the order of energy minima slightly in the current DFT study. The overall agreement between the DFT and MP2 free energies is reasonable. Both methods give the C_5 conformer as the global minimum, with the β conformer with slightly higher free energy (0.39 for MP2 and 0.06 kcal/mol for DFT). Two other conformers, α_R and β_2, follow C_5-β with relative free energies roughly 1−1.6 kcal/mol higher than that of the C_5. Both methods partially agree with the experimental observations that the β and α_R conformers are the dominant species in water.[27,32,37,38,40] However, the lack of experimental evidence on the existence of conformers C_5 and β_2 does not explain the theoretical results in which they have very similar energies with β and α_R, respectively. As indicated by Wang et al., the lack of explicit interactions of AD with water molecules as a result of using the PCM model may attribute to this disagreement. They have also speculated two intermediate conformers with explicit hydrogen-bonded water molecules whose geometries lie between C_5 and β and α_R and β_2, respectively. These intermediate conformers are believed to be the dominant species β and α_R observed in experiments.

The differences in the relative free energies between MP2 and DFT are similar to the differences in the relative energies before the free energy corrections, indicating inheritance of the problem. That is, there is a general trend of more stable extended structures and less stable compact structures from the DFT method than the MP2 methods. Another significant discrepancy is that the free energy (relative to C_5) of the α_L conformer in the DFT results is about 1.63 kcal/mol higher than that from the MP2 results. This energy difference comes from energy without corrections (1.57). There is no direct

experimental or theoretical evidence to determine which set of results is closer to the real value. However, from the MP2 results, the free energy of α_L is just 0.7 kcal/mol higher than the α_R conformer. This small difference should have resulted in a population of the α_L conformer being observed experimentally. Therefore, a much larger energy gap of 1.7 kcal/mol between the α_L and α_R conformers from the DFT calculations fits better with the experimental results where only the population of α_R and no α_L is observed.[27,37]

The free energy corrections for the aqueous energy map have been subjected to the aforementioned linear smoothing technique. Analysis also has shown that the systematic errors caused by constraints of dihedral angles in optimization are similar to those of the gas-phase free energy corrections in terms of behaviors and magnitudes. As shown in Figure 2D, the aqueous FES is slightly nosier than that of the gas-phase FES. This is most likely because the convergence criteria used for partial geometry optimization of AD in aqueous phase at each grid point is slightly larger than that used for gas-phase optimization. The remaining forces will lead to larger noises in frequency calculations and therefore thermal corrections. Similar to the gas-phase behavior, the free energy corrections tend to raise the slope of the energy profiles. The increase is even greater in the aqueous phase than in the gas phase. The average free energy correction to the transition state is 1.1 kcal/mol higher than the average to the minima, as comparing to 0.7 kcal/mol in gas phase.

In a brief summary, the current B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) with PCM model approach was able to generate energy contour maps of AD in the gas phase and in water that is at least as accurate as the previous MP2 approaches. The cost-efficient hybrid DFT method has also allowed us to generate the energy contour maps using a finer grid as well as performing frequency calculation at each grid point to generate free energy corrections to the map. The validated ab initio FESs will be used as standards to access the accuracy of FESs generated by metadynamics simulations in combination with several commonly used force fields, namely AMBER94, AMBER03, CHARMM27, and OPLSAA.

Free Energy Surfaces of AD by Metadynamics Simulations using MM Force Fields. Figure 3 displays FESs of AD in a gas phase, and Figure 4 exhibits FESs of AD in an aqueous solution. In this section, we will visually examine and compare the general features of the force field FESs against that of the ab initio FESs to provide a qualitatively assessment of different force fields and the metadynamics method.

The gas-phase FESs are obtained using both hills setting (1) (Figure 3A) and (2) (Figure S1, Supporting Information). Comparison of the gas-phase FESs from two hills setting shows that using the same force field but different hill sizes produces almost identical FESs for each force field. This indicates the robustness of the metadynamics method to hills settings. The efficiency of the metadynamics simulations in sampling conformational space of the selected variables is shown in Figure 3, in which the contour lines cover up almost all conformational space except a very small high-energy area at around $\phi = 0°$ and $\psi = 180°$. The excellent coverage indicates a good conformational sampling in just 5 ns using hills setting (1). The dependency of the sampling efficiency on the metadynamics (hills) parameters can be analyzed based on the difference in conformational coverage. The metadynamics simulations using hills setting (2) add Gaussian potentials (hills) with half of the width and less frequently when compared with hills setting (1). Therefore, even with a 20 ns of simulation, the uncovered area is still larger than that of the FES from hills setting (1) with 5 ns. However, the blank areas are high-energy hills and generally do not affect the minima and important transition states. The unsampled areas in the aqueous solution phase FESs (Figure 4) are much less than that of the gas phase (using hills setting (2)). The result indicates that solvation effect has made the free energy surface much more accessible, which is consistent with what we observed in the ab initio calculations.

The general shapes of all force field gas-phase FESs are similar to that of the ab initio FES (Figure 2B). All force field FESs capture the major minima C_5, $C7_{eq}$, and $C7_{ax}$. Their positions in terms of the values of (ϕ, ψ) are also in good agreement with that of the ab initio FES. However, further inspection of Figure 3 indicates that there are characteristic differences in details of the gas-phase FESs of different force fields and the ab initio FES. For example, most force field FESs do not have local minima for conformers α_L and α_D, whose positions are clearly defined in the ab initio FES. On the OPLSAA, OPLSAA/L, or the AMBER94 FES, there is a small relatively flat region below the $C7_{ax}$ minimum that can be classified as the α_D conformer. For α_L, a similar flat region above $C7_{ax}$ is seen only on the OPLSAA FES. Second, on the AMBER03, OPLSAA, and CHARMM27 force field FESs, the stability of the β_2 conformer area is underestimated. It should be noted that Mackerell et al. have noticed this underestimation and have added ϕ, ψ dihedral cross terms (CMAP) to the CHARMM force field to improve the energetics of the β_2 conformer.[41] However, the current FES is calculated without the cross terms to evaluate the force fields within the boundary of the standard functional forms. Another significant difference between the force field and ab initio FES is that the CHARMM27 force field has overestimated the stabilities of the conformational space at the lower left quadrant of the FES, i.e., the area around (−120°, −120°).

The rest of the transition states fall into two major groups that connect the low-energy regions (LeRI and LeRII, C_5-$C7_{eq}$-β-β_2-α_R) on the negative ϕ side, to another region on the positive ϕ side (LeRIII, α_L-$C7_{ax}$-α_D) through rotation of the dihedral angle ϕ clockwise and counterclockwise, respectively. One group includes TS1−TS3 which have ϕ roughly at 0°. As shown in Figure 3, only the AMBER94 force field FES features all three transition states with TS3 shifted slightly upward (larger ψ). The OPLSAA and CHARMM27 FESs feature TS1 and another transition state (TS2/TS3) which positions somewhere between the ab initio TS2 and TS3. The AMBER03 and OPLSAA/L force fields, on the other hand, only show TS2, which also shifted slightly downward (smaller ψ) in the AMBER03 FES. The other group includes TS4 and TS5 whose ϕ angle values are roughly at 120°−150°. Both transition states are featured in the FESs of three force fields, AMBER94, OPLSAA, and

OPLSAA/L, while only TS5 is shown by AMBER03, and a combined TS4/TS5 is shown by CHARMM27. The last transition state TS6 is of less importance due to its higher free energy (∼11 kcal/mol relative to $C7_{eq}$ by DFT). Generally, all force fields have predicted its position reasonably well but overestimated the energy barrier with respect to the global mimimum $C7_{eq}$, in which OPLSAA and OPLSAA/L did slightly better than the other three force fields.

The comparison of Figures 3 and 4 shows that the changes from the gas phase to the solution phase FESs are common among all force fields and also very similar to those observed in the ab initio FESs (Figure 2). These changes include more extended low-energy regions, flatter surface, diminishing of gas-phase global minimum $C7_{eq}$ and emerging of β and α_R, and shifting from $C7_{ax}$ to α_L. Considering the force field aqueous FESs are obtained from simulations with explicit solvent molecules, the results indicate that the PCM model used in the ab initio calculations, although lacking specific AD-water interactions, does capture the major solvation effect.

On the other hand, the difference in the details of the FESs between force field and ab initio is also more evident in the aqueous solution phase. Although all force fields have captured the low-lying minima β and α_R accurately in terms of both their energies and positions. When compared with the ab initio FES, AMBER94, and CHARMM27 fail to capture minimum C_5. For the β_2 conformer, OPLSAA/L has done a good job, followed by the two AMBER force fields which predict the position of β_2 to be slightly higher than that of the ab initio calculations. The OPLSAA gives a local minimum roughly 60° (in ψ) above the position of the ab initio β_2 conformer. The CHARMM27 force field has failed to produce a β_2 minimum and underestimated the stability of the β_2 region significantly. The performance of the force fields on the second group of minima α_L-$C7_{ax}$-α_D-α_D' on the positive (in ϕ) side of the contour map also varies. The OPLSAA force field seems to produce the best fit to the ab initio result, featuring well-separated α_L and $C7_{ax}$ minima with α_L being the lowest in free energy among the four minima. The two AMBER and the OPLSAA force fields combine α_L and $C7_{ax}$ into one minimum, which positions in the middle. The CHARMM27 force field predicts the position of α_L quite well but underestimates its stability significantly. In terms of transition states, all force fields except CHARMM27 predict the position of TS7 fairly well, and its stability varies from 2 to 6 kcal/mol with respect to the α_R or β conformer. Similar to the gas-phase CHARMM27 FES, the stability of the lower left quadrant of the aqueous CHARMM27 FES has been overestimated, therefore, no TS7 is observed. Instead, a transition state above the α_R and below the $C7_{eq}$ has been identified. For the transition-state group TS1−TS3, all force field FESs feature TS1 and TS3 whose positions and energetics also agree well with the ab initio FES. TS2 is missing from all force field energy maps, however, it is not of serious concern since TS2 is much higher in energy than the other two transition states. All force field FESs feature both TS5 and TS8, and their positions similar to that have been observed in the ab initio FES.

In addition to the force field-specific discrepancies discussed above, there is also one common difference between all MM force field FESs and the ab initio FES of AD in aqueous phase, i.e., the free energy span of the ab initio FES (0 to ∼14 kcal/mol), which is wider than that of the force field FESs (0 to ∼10−12 kcal/mol). Since we have excluded the high-energy areas of the FES when calculating the above free energy range, we can safely say that the insufficient conformational sampling in these high-energy areas by the force fields does not influence our results. The lack of specific interactions (hydrogen bonds) between solvent and AD in the ab initio PCM calculations may have contributed to this discrepancy. We could not determine to what extent the lack of specific interactions has on the FES until ab initio MD simulation of AD with explicit solvent can be afforded. In addition, the free energy corrections may also have a small contribution (0.2−0.4 kcal/mol) to this common discrepancy.

The qualitative visual inspection of both ab initio and force field FESs reveals that although they all bear similar general features, they are distinctly different in characteristic details. In the following section, we will present quantitative assessment of the performance of each force field.

Quantitative Assessment of the Force Field FESs. Before a detailed quantitative comparison between force field and ab initio FESs can be made, it is necessary to determine the error (noise) of each individual FES so that the difference in energy measured later can be put into context. The error of the ab initio FES, ignoring errors from potential energy which is not the scope of this paper, is 0.2−0.4 kcal/mol based on analysis on the free energy corrections in the previous section. For the force field FESs, the method we use to estimate their resolutions was briefly mentioned in a previous article.[51] Basically, the errors are calculated as the standard deviation of the FEDM between two FESs that should be the same, e.g., two FESs using the same force field but different hills settings. The errors are estimated to be around 0.3−0.4 kcal/mol, or slightly larger (up to 0.7 kcal/mol) in some areas due to insufficient sampling, for both gas-phase and aqueous phase FESs. More details of the error analysis can be found in the Supporting Information.

As we have mentioned in the Introduction, most force field parametrization processes that integrate ab initio calculations have focused on getting the position and the relative energies of a few low-lying minima right. Since such parametrization does not pay much attention to the transition states, the resulting force field could introduce large errors when studying protein dynamics and kinetics. In this paper, we present a strategy to assess the performance of any force field not only in the low-energy regions but also transition-state areas. As shown in Figure 5, the $\phi-\psi$ space has been partitioned into six regions, which in total account for more than 80% of the conformational space. The rest of the space is of high energy and excluded from our analysis. The partitions are different from gas to aqueous solution phase and are based on common practice of protein backbone conformational partitions as well as the ab initio FESs. There are three low-energy regions in which LeRI encloses minima C_5, β, and $C7_{eq}$, and LeRII includes minima β_2 and α_R, and both of these regions are on the negative ϕ side. LeRIII is on the

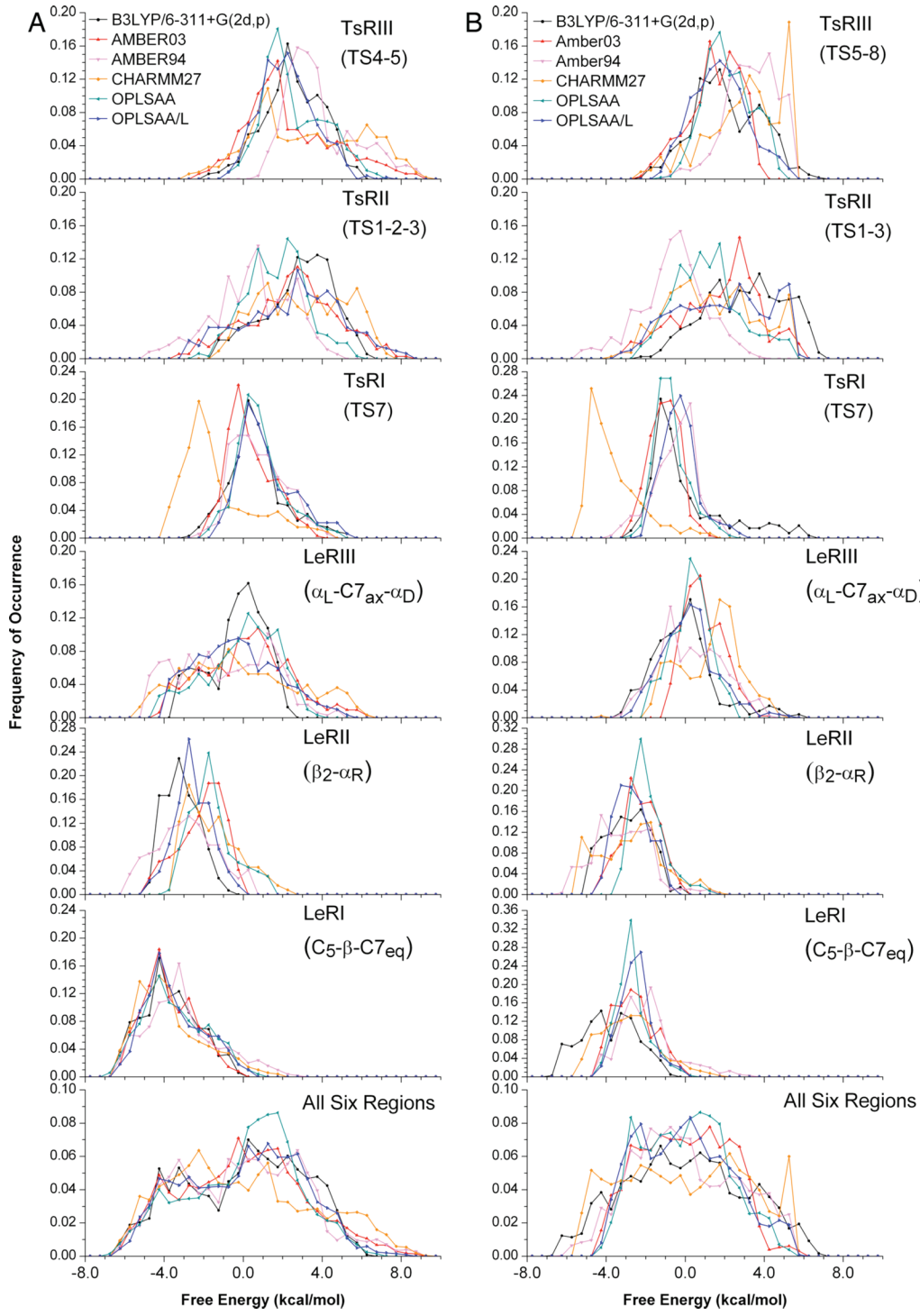


Figure 6. Free energy distributions: (A) gas and (B) aqueous solution phases.

positive ϕ side and covers minima α_L, $C7_{ax}$, and α_D. There are also three transition regions in which TsRI is the low-lying transition region between LeRII and LeRI. Both TsRII and TsRIII link the negative and positive ϕ sides through rotation of the ϕ dihedral angle clockwise (through 0°) or counterclockwise (through 180°).

Based on the 50 × 50 grid map, the distribution of free energies of each region has been plotted in Figure 6 for each force field or ab initio FES in gas or aqueous solution phase. In addition, we have also measured two quantities per region for each force field FES. As shown in Tables 3 and 4, the first one is the mean free energy, and the other is the standard deviation of the FEDM between the force field and ab initio FESs for each region. Unlike the FESs (Figures 2−4), where free energy is relative to the global minimum, the mean free energies in Tables 3 and 4 or the free energies in the distribution plots of Figure 6 are relative to the overall mean of all six regions for each force field. This overall mean, as a reference, is unbiased and not subject to the error in energy of a single point (the global minimum) on the FES. Comparison of the mean free energy of each region between the ab initio and any force field FESs roughly indicates the average accuracy of the force field on energetics of a particular group of conformers or transition states. The

Table 3. Mean Free Energies of the Six Regions of the Ab Initio and MM Force Field FESs

	DFT	AMBER03	AMBER94	CHARMM27	OPLSAA	OPLSAA/L
			Gas Phase[a]			
all six regions	0.00	0.00	0.00	0.00	0.00	0.00
LeRI (C_5-β-$C7_{eq}$)	−3.72	−3.82/−3.74	−3.16/−3.07	−4.14/−3.69	−3.69/−3.50	−3.63/−3.51
LeRII (β_2-α_R)	−3.11	−1.86/−2.09	−2.60/−2.89	−1.27/−1.55	−1.22/−1.60	−2.66/−2.57
LeRIII (α_L-$C7_{ax}$-α_D)	−0.32	0.42/0.18	−0.35/−0.88	0.90/−0.11	0.32/−0.07	−0.07/−0.31
TSRI (TS7)	0.67	0.45/0.35	0.64/0.76	−1.51/−1.29	0.81/0.85	1.08/1.15
TSRII (TS1−TS3)	2.82	2.82/2.50	0.74/0.48	3.02/3.01	2.10/1.70	2.27/2.49
TSRIII (TS4−TS5)	2.48	1.87/2.34	3.38/3.86	2.87/3.28	2.11/2.54	2.26/2.23
			Aqueous Solution			
all six regions	0.00	0.00	0.00	0.00	0.00	0.00
LeRI (C_5-β-$C7_{eq}$)	−3.85	−2.57	−1.91	−2.79	−2.68	−2.52
LeRII (β_2-α_R)	−2.90	−2.29	−3.21	−2.74	−1.90	−2.66
LeRIII (α_L-$C7_{ax}$-α_D)	−0.07	0.94	0.30	1.27	0.26	0.21
TSRI (TS7)	0.04	−1.03	−0.50	−3.60	−0.71	−0.27
TSRII (TS1−TS3)	3.02	1.74	−0.60	1.37	1.12	1.87
TSRIII (TS5−TS8)	2.07	1.48	3.34	2.97	2.15	1.75

[a] First number is from a 20 ns gas-phase simulation using hills setting (2), while second number after the slash is from a 5 ns gas-phase simulation using hills setting (1).

Table 4. Standard Deviation (kcal/mol) of the FEDMs between the Ab Initio and Force Field FESs

	AMBER03	AMBER94	CHARMM27	OPLSAA	OPLSAA/L
		Gas Phase[a]			
all six regions	1.50/1.39	1.78/1.96	2.13/2.17	1.31/1.30	1.25/1.25
LeRI (C_5-β-$C7_{eq}$)	0.89/0.75	1.39/1.53	1.03/1.06	0.76/0.89	0.89/0.89
LeRII (β_2-α_R)	1.09/1.24	1.25/1.45	1.03/1.21	1.04/1.07	1.03/0.99
LeRIII (α_L-$C7_{ax}$-α_D)	1.58/1.53	1.92/1.83	2.21/2.32	1.03/1.09	1.37/1.51
TSRI (TS7)	1.01/0.90	1.04/1.15	2.25/2.22	1.04/0.98	0.76/0.73
TSRII (TS1−TS3)	1.55/1.55	1.64/1.62	1.65/2.02	1.47/1.60	1.96/1.90
TSRIII (TS4−TS5)	1.82/1.70	1.34/1.45	2.14/2.28	1.41/1.24	0.87/1.00
		Aqueous Solution			
all six regions	1.67	2.50	2.52	1.62	1.51
LeRI (C_5-β-$C7_{eq}$)	1.23	1.84	1.69	1.26	1.17
LeRII (β_2-α_R)	1.12	1.45	1.87	1.15	0.79
LeRIII (α_L-$C7_{ax}$-α_D)	1.23	1.96	2.11	1.27	1.19
TSRI (TS7)	1.74	2.17	2.14	1.55	1.32
TSRII (TS1−TS3)	1.50	1.52	1.92	1.31	1.82
TSRIII (TS5−TS8)	1.30	1.39	1.94	1.05	1.22

[a] First number is from a 20 ns gas-phase simulation using hills setting (2), while second number after the slash is from a 5 ns gas-phase simulation using hills setting (1).

standard deviation of the FEDM indicates how well the two energy surfaces match, therefore is a good measure of whether or not the force field predicts the positions of minima or transition states as accurately as the ab initio method. We will discuss, region by region, first the results of the gas-phase FESs, then the aqueous phase FESs.

Gas-Phase Region by Region. As shown in the bottom panel of Figure 6A, the overall free energy distributions, obtained by metadynamics simulations using the five different force fields, fit reasonably well with the ab initio calculations. The width and the position of the peaks all match closely with the ab initio distribution. Table 4 shows that the standard deviation of the difference between any force field and the ab initio FESs ranges from 1.25 to 2.17 kcal/mol, which is much larger than the estimated errors (0.2−0.7 kcal/mol), indicating true difference. More specifically, the two OPLSAA and the AMBER03 force fields differ from the ab initio free energy map about 1.2−1.4 kcal/mol, while AMBER94 and CHARMM27 have larger deviations from 1.8 to 2.2 kcal/mol.

The performance of all force fields in the first low-energy region, LeRI, is also good. As shown in Table 3, the average free energies (−3.07 and −3.50 to −3.74) of LeRI are mostly within the error range with that of the ab initio FES (−3.72). The only exception is AMBER94 (−3.07), which underestimates the stability of LeRI for about 0.65 kcal/mol, just outside the error range of 0.3−0.4 kcal/mol. The standard deviation of the difference energy map between any force field and ab initio FES in this region mostly ranges from 0.8−1.0 kcal/mol. The AMBER94, again, has a slightly larger deviation of 1.5 kcal/mol from the ab initio energy map. In gas phase, all five force fields seem to underestimate the stability of the second low-energy region LeRII, i.e., the β_2-α_R conformers, especially the CHARMM27, OPLSAA, and AMBER03 force fields, whose average free energies are 1.0−1.5 kcal/mol higher than the ab initio mean of LeRII. Both AMBER94 and OPLSAA/L give reasonable average free energies. However, the distribution of free energies (as shown in Figure 6A) of AMBER94 is much wider, and not surprisingly the standard deviation from the ab initio map is the largest (1.5 kcal/mol). That leaves the OPLSAA/L force field that performs well in both average free energy (∼0.5 kcal/mol difference from ab initio) and standard deviation (∼1 kcal/mol). For the third low-energy region LeRIII, the

average free energies (−0.88 to 0.28) of all five force fields are all within a reasonable range, comparing with the ab initio mean (−0.32) of LeRIII. Figure 6A shows that the force field free energies distributions are usually wider than that of the ab initio. The standard deviations of the difference between any force field and the ab initio FESs rank as the following: OPLSAA has the smallest deviation (1.09 kcal/mol), followed by OPLSAA/L and AMBER03 (∼1.5 kcal/mol), then followed by AMBER94 and CHARMM27 (1.8−2.3 kcal/mol). This is consistent with the observations based on the FESs, in which most force field FESs miss the α_L and α_D conformers, except OPLSAA which shows some signs of the two minima.

The next region, TsRI, features transition state TS7 and minimum α′. The performance of all force fields except CHARMM27 are good, with their mean free energies within ±0.4 kcal/mol, standard deviation ∼1.0 kcal/mol, and distribution fits well with the ab initio calculations. The exception, CHARMM27, heavily overestimates the stability of this region. The mean free energy calculated by CHARMM27 is nearly 2 kcal/mol lower than the ab initio mean. In addition, the standard deviation of the difference map between the CHARMM27 and ab initio FESs in this region is 2.2 kcal/mol, which is two times of any other force field.

The second transition-state region, TsRII, is one of the two paths between the $\phi < 0$ (LeRI and LeRII) and the $\phi > 0$ (LeRIII) regions. We have observed a general increase in standard deviation from the minima regions to this transition region. The results indicate that match between any force field FES with the ab initio FES in this region is less satisfactory than that of the low-energy minima regions. This is probably due to the fact that all parametrization focuses only on matching the positions and energetics of minima. The best standard deviation is ∼1.5 kcal/mol from the two AMBER and the OPLSAA force fields, while the other two are at ∼2.0 kcal/mol. The average free energies of TsRII of AMBER03, OPLSAA/L, and CHARMM27 are reasonable, while OPLSAA and AMBER94 underestimate the average free energy by 1.1−2.3 kcal/mol, respectively. Overall, the AMBER03 force field gives the best fit to the ab initio FES in this region (TsRII).

The last transition-state region, TsRIII, also links the $\phi < 0$ (LeRI and LeRII) and the $\phi > 0$ (LeRIII) regions but from an opposite direction. AMBER94 and CHARMM27 overestimate the energy of this region for 1.4 and 0.8 kcal/mol, respectively. The average free energy of the other three force fields is good, however, the standard deviation between AMBER03 (1.7 kcal/mol) and ab initio is larger than the two OPLSAA (1.0 to 1.2 kcal/mol) force fields. Overall, the OPLSAA/L FES fits the best with ab initio for TsRIII, in terms of both mean and standard deviation.

Aqueous Phase Region by Region. Consistent with what we have observed in the free energy maps, the distribution of free energies among all six regions is generally wider for ab initio than the force fields. It should be noted that the sharp peaks we saw at the high energy end of the CHARMM27 distribution are due to insufficient conformational sampling, mostly in TsRIII region, as shown in Figure 4. Standard deviations between the force field and ab initio FESs are generally larger than that in the gas phase, indicating less satisfactory matches. As shown in Table 4, the standard deviation of individual region or among all six regions falls into two groups. The OPLSAA, OPLSAA/L, and AMBER03 force fields are in one group (1.5−1.7 kcal/mol) with smaller deviations than AMBER94 and CHARMM27 (2.5 kcal/mol). This trend is similar to the performance of these force fields in gas phase.

For the first low-energy region, LeRI (C_5-β-$C7_{eq}$), the average free energies of all force field FESs are higher than the ab initio mean. This is directly related to the fact that ab initio has a wider distribution, therefore the average free energy of the lowest energy region, relative to the average of the overall distribution, is lower than that of the force fields. Since energy is only relevant in terms of relative stability, we will discuss the effects of this discrepancy later together with the average energies of other regions. However, we would like to mention that the average free energy of LeRI of AMBER94 is 2.0 kcal/mol higher than the ab initio mean. This difference is slightly larger than the 1.0−1.3 kcal/mol differences between the other force fields and ab initio FESs. This is consistent with the gas-phase results in which AMBER94 underestimates the stability of LeRI. In terms of standard deviation, i.e., how well a force field FES matches with the ab initio FES, OPLSAA/L, OPLSAA, and AMBER03 give ∼1.2 kcal/mol deviation, while AMBER94 and CHARMM27 go up to 1.7−1.8 kcal/mol.

For the second low-energy region, LeRII (β_2-α_R), the average free energy is 0.9 kcal/mol higher than that of LeRI (C_5-β-$C7_{eq}$) based on the ab initio FES. This difference is very much in agreement with a probability partition of 80−20 between the C_5-β (or β-P_{II}) and the α_R states, according to an experimental/computational investigation.[27] For AMBER03, CHARMM27, and OPLSAA/L, the mean free energy of LeRII agrees well with that of the ab initio. However, the free energy differences between the LeRI and LeRII regions for these force fields are too small (−0.3, −0.05, and 0.1 kcal/mol, respectively) because of the underestimation of the stability of the LeRI regions. The near to zero (−0.05) energy difference obtained by CHARMM27 is consistent with previous MD simulation which gives a 50−50 probability distribution[24] between the two regions. AMBER03 is slightly better than CHARMM27, while OPLSAA/L is slightly worse. For AMBER94, the mean free energy of LeRII is in line with ab initio as well. However, the order of mean free energy between LeRI and LeRII is reversed due to the underestimation of stability of LeRI. Again, the reversed order of stability is consistent with previous MD simulation, which gives a reversed 20−80 probability distribution between LeRI and LeRII.[24] For OPLSAA, the mean free energy of LeRII is 1.0 kcal/mol higher than the ab initio mean. However, since both LeRI and LeRII's stabilities have been underestimated by 1.0−1.1 kcal/mol, their relative stability is similar to ab initio. This is again in line with previous MD simulation that gives an 85−15 probability distribution between the two regions.[24] The consistency observed also validates the current approach of using the mean free energy and the definition of regions. The standard deviations for LeRII between the force field

and ab initio FESs are 0.8 kcal/mol for OPLSAA/L, ~1.1 kcal/mol for OPLSAA and AMBER03, and 1.5−1.9 kcal/mol for AMBER94 and CHARMM27.

For the low-energy region on the negative ϕ side, LeRIII, the mean free energies are 1.0−1.3 kcal/mol higher than the ab initio mean for AMBER03 and CHARMM27. This results in relative stability between LeRI and LeRIII, in good agreement with ab initio, but not for relative stability between LeRII and LeRIII. For the other three force fields, the mean free energies of LeRIII are close to the ab initio value. Therefore, the relative stabilities between LeRII and LeRIII are good for AMBER94 and OPLSAA/L but not so good for OPLSAA due to the higher mean free energy of LeRII. Furthermore, none of these three produces a good relative stability between LeRI and LeRIII. The standard deviations of LeRIII between force field and ab initio are very similar to LeRI, in which the first group of force fields OPLSAA/L, OPLSAA, and AMBER03 is at ~1.2 kcal/mol, while the other group including AMBER94 and CHARMM27 is at ~2.0 kcal/mol.

For the first transition-state region, TSRI, the results are very similar to the gas phase, in which CHARMM27 heavily and AMBER03 and OPLSAA slightly overestimate the stability of this area. Combining performance in both mean and standard deviation, the OPLSAA/L is the best as comparing with ab initio. For the second transition-state region, TSRII, the average free energy from all force fields is lower than that of the ab initio. This directly contributes to the fact that ab initio FES has a wider free energy distribution than all the force field FESs. More specifically, AMBER94 underestimates the free energy of this region for 3.6 kcal/mol. The other force fields perform better than AMBER94 and only underestimate the average free energy for 1.1 to 1.9 kcal/mol. Similar to the gas-phase result, AMBER03 is the best match to the ab initio values combining the performance of mean and standard deviation. OPLSAA/L has the best mean but slightly larger deviation, while OPLSAA has the best deviation but slightly lower mean. For the third transition-state region, TsRIII, the mean free energies obtained by CHARMM27 and AMBER94 are 0.9 and 1.3 kcal/mol higher than the ab initio mean. Both force fields also have higher standard deviations from the ab initio FES. The other three force fields match with the ab initio values in both mean free energy (difference within 0.6 kcal/mol) and standard deviation (1.0−1.3 kcal/mol). The two OPLSAA force fields perform slightly better than AMBER03 in both mean and standard deviation.

Conclusions

The current B3LYP/6-311+G(2d,p)//B3LYP/6-31G(d,p) with PCM model approach is able to generate potential energy contour maps of AD in gas phase and in water that is at least as accurate as the previous MP2 approaches. The cost-efficient hybrid DFT method has also allowed us to generate FESs by calculating free energy corrections on a 50 × 50 grid base. The error range, from the free energy corrections, is estimated at around 0.2−0.4 kcal/mol. The average free energy difference of 0.9 kcal/mol between the first two conformational basins, LeRI (C_5-β-$C7_{eq}$) and LeRII (β_2-α_R), on the ab initio-DFT FES in aqueous phase agrees excellently with a 80−20 probability distribution from recent experimental/computational investigations.[27,38]

Metadynamics simulations, in combination with five commonly used MM force fields, have been carried out to obtain FESs of AD in both gas phase and water. The error range of these FESs is calculated to be 0.3−0.4 kcal/mol and in some conformational areas goes up to 0.7 kcal/mol due to sampling. Quantitative assessment of these force field FESs in three low-energy conformational basins, LeRI (C_5-β-$C7_{eq}$), LeRII (β_2-α_R), and LeRIII(α_L-$C7_{ax}$-α_D) as well as three transition-state regions was made according to the ab initio DFT FESs. The average free energy differences between the LeRI and LeRII basins, among several force fields, agree well with previous MD simulations,[24] which gives validation to the partition of conformational basin and the assessment method. The quantitative assessment reveals variations in performance from one conformational region to another and from force field to force field or from gas to aqueous phase. Although not one MM force field is able to outperform all others in all conformational areas, the overall best performer is the OPLSAA/L force field, followed by OPLSAA and AMBER03. The results also indicate a certain degree of transferability of performance from gas to aqueous phase. However, there are also areas where a better fit in one phase leads to a decrease performance of another phase, such as AMBER03 or OPLSAA/L.

In summary, we have presented a new method of assessing force field performance not only in the energies of conformers but also transition-state barriers. The method presented and results obtained here should be useful for improving the parametrization of force field. More specifically, the ab initio-DFT FESs, the partitions of conformational basins, and the transition-state regions as well as the method of quantitative assessment can be used in force field parametrization to improve the force field description of all areas of the ($\phi-\psi$) conformational map.

Acknowledgment. Z.L. and P.B.M. acknowledge the financial support from the National Science Foundation (EMT CCF-0622162 and MRI CHE-0420556), the National Health Institutes (GM075990) andthe H. O. West Foundation. B.E. acknowledges financial support through a VIDI grant from the Chemical Sciences Division of The Netherlands Organization for Scientific Research (NWO−CW). We also thank the TeraGrid Project for providing computational resources.

Supporting Information Available: Details on the series of partial optimization to obtain the α_D conformer and the TS3 transition state for gas phase and the β conformer for aquesous phase. Details on the error analysis of the force field FESs. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U.K., 1987.

(2) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press: San Diego, CA, 1996.

(3) Cheatham, T. E., III; Kollman, P. A. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435–471.

(4) Leach, A. R. *Molecular Modelling: Principles and Applications*. Prentice Hall: Upper Saddle River, NJ, 2000.

(5) Becker, O. M.; MacKerell, A. D., Jr.; Roux, B.; Watanabe, M. *Computational Biochemistry and Biophysics*, 1st ed.; Marcel Dekker: New York, 2001.

(6) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211–243.

(7) Becker, O. M.; Karplus, M. *Guide to Biomolecular Simulations*. Springer: Dordrecht, The Netherlands, 2005.

(8) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.

(9) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(10) Schuler, L. D.; Daura, X.; van, G. W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22* (11), 1205–1218.

(11) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(12) Anderson, A. G.; Hermans, J. Microfolding - conformational probability map for the alanine dipeptide in water from molecular-dynamics simulations. *Proteins* **1988**, *3* (4), 262–265.

(13) Wu, X. W.; Sung, S. S. Simulation of peptide folding with explicit water--a mean solvation method. *Proteins* **1999**, *34* (3), 295–302.

(14) Brooks, C. L., III. Protein and peptide folding explored with molecular simulations. *Acc. Chem. Res.* **2002**, *35* (6), 447–454.

(15) Zhou, R. Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins* **2003**, *53* (2), 148–161.

(16) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β-hairpin peptide. *J. Phys. Chem. B* **2004**, *108* (21), 6582–6594.

(17) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.

(18) Lei, H.; Duan, Y. Protein folding and unfolding by all-atom molecular dynamics simulations. *Methods Mol. Biol.* **2008**, *443*, 277–295.

(19) van, d. K. M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A. C.; Daggett, V. Dynameomics: A comprehensive database of protein dynamics. *Structure* **2010**, *18* (4), 423–435.

(20) Velez-Vega, C.; Borrero, E. E.; Escobedo, F. A. Kinetics and reaction coordinate for the isomerization of alanine dipeptide by a forward flux sampling protocol. *J. Chem. Phys.* **2009**, *130* (22), 225101–225112.

(21) Gaigeot, M. P. Unraveling the conformational dynamics of the aqueous alanine dipeptide with first-principle molecular dynamics. *J. Phys. Chem. B* **2009**, *113* (30), 10059–10062.

(22) Han, W.-G.; Jalkanen, K. J.; Elstner, M.; Suhai, S. Theoretical study of aqueous N-Acetyl-L-alanine N'-Methylamide: Structures and raman, VCD, and ROA spectra. *J. Phys. Chem. B* **1998**, *102* (14), 2587–2602.

(23) Smith, P. E. The alanine dipeptide free energy surface in solution. *J. Chem. Phys.* **1999**, *111* (12), 5568–5579.

(24) Hu, H.; Elstner, M.; Hermans, J. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins* **2003**, *50* (3), 451–463.

(25) Wang, Z.-X.; Duan, Y. Solvation effects on alanine dipeptide: A MP2/cc-pVTZ//MP2/6-31G** study of (ϕ, ψ) energy maps and conformers in the gas phase, ether, and water. *J. Comput. Chem.* **2004**, *25* (14), 1699–1716.

(26) Feig, M. Kinetics from implicit solvent simulations of biomolecules as a function of viscosity. *J. Chem. Theory Comput.* **2007**, *3* (5), 1734–1748.

(27) Kwac, K.; Lee, K.-K.; Han, J. B.; Oh, K.-I.; Cho, M. Classical and quantum mechanical/molecular mechanical molecular dynamics simulations of alanine dipeptide in water: Comparisons with IR and vibrational circular dichroism spectra. *J. Chem. Phys.* **2008**, *128* (10), 105106–105119.

(28) Liu, C.; Zhao, D.-X.; Yang, Z.-Z. ABEEM$\sigma\pi$ fluctuating charge force field applied to alanine dipeptide and alanine dipeptide-water systems. *J. Theor. Comput. Chem.* **2010**, *9* (Supp. 1), 77–97.

(29) Jono, R.; Watanabe, Y.; Shimizu, K.; Terada, T. Multicanonical *ab inito* QM/MM molecular dynamics simulation of a peptide in an aqueous environment. *J. Comput. Chem.* **2010**, *31* (6), 1168–1175.

(30) Weise, C. F.; Weisshaar, J. C. Conformational analysis of alanine dipeptide from dipolar couplings in a water-based liquid crystal. *J. Phys. Chem. B* **2003**, *107* (14), 3265–3277.

(31) Lavrich, R. J.; Plusquellic, D. F.; Suenram, R. D.; Fraser, G. T.; Hight, W. A. R.; Tubergen, M. J. Experimental studies of peptide bonds: Identification of the C7eq conformation of the alanine dipeptide analog N-acetylalanine N'-methylamide from torsion-rotation interactions. *J. Chem. Phys.* **2003**, *118* (3), 1253–1265.

(32) Madison, V.; Kopple, K. D. Solvent-dependent conformational distributions of some dipeptides. *J. Am. Chem. Soc.* **1980**, *102* (15), 4855–63.

(33) Deng, Z.; Polavarapu, P. L.; Ford, S. J.; Hecht, L.; Barron, L. D.; Ewig, C. S.; Jalkanen, K. Solution-phase conformations of N-Acetyl-N'-methyl-L-alaninamide from vibrational raman optical activity. *J. Phys. Chem.* **1996**, *100* (6), 2025–34.

(34) Poon, C.-D.; Samulski, E. T.; Weise, C. F.; Weisshaar, J. C. Do bridging water molecules dictate the structure of a model dipeptide in aqueous solution. *J. Am. Chem. Soc.* **2000**, *122* (23), 5642–5643.

(35) Takekiyo, T.; Imai, T.; Kato, M.; Taniguchi, Y. Temperature and pressure effects on conformational equilibria of alanine dipeptide in aqueous solution. *Biopolymers* **2004**, *73* (2), 283–290.

(36) Bohr, H. G.; Jalkanen, K. J. Spectroscopic studies of small biomolecules in condensed phases. *Condens. Matter Theor.* **2006**, *20*, 375–391.

(37) Kim, Y. S.; Wang, J.; Hochstrasser, R. M. Two-dimensional infrared spectroscopy of the alanine dipeptide in aqueous solution. *J. Phys. Chem. B* **2005**, *109* (15), 7511–7521.

(38) Grdadolnik, J.; Golic, G. S.; Avbelj, F. Determination of conformational preferences of dipeptides using vibrational spectroscopy. *J. Phys. Chem. B* **2008**, *112* (9), 2712–2718.

(39) Mukhopadhyay, P.; Zuber, G.; Beratan, D. N. Characterizing aqueous solution conformations of a peptide backbone using Raman optical activity computations. *Biophys. J.* **2008**, *95* (12), 5574–5586.

(40) Schweitzer-Stenner, R.; Measey, T.; Kakalis, L.; Jordan, F.; Pizzanelli, S.; Forte, C.; Griebenow, K. Conformations of alanine-based peptides in water probed by FTIR, raman, vibrational circular dichroism, electronic circular dichroism, and NMR spectroscopy. *Biochemistry* **2007**, *46* (6), 1587–1596.

(41) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L. III Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400–1415.

(42) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24* (16), 1999–2012.

(43) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.

(44) Chipot, C.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer-Verlag: Berlin, Germany, 2007.

(45) Smiatek, J.; Heuer, A. Calculation of free energy landscapes: a histogram reweighted metadynamics approach. *Physics* **2010**, 1–11; arXiv:1006.4308v1 [physics.comp-ph].

(46) Dickson, B. M.; Legoll, F.; Lelievre, T.; Stoltz, G.; Fleurat-Lessard, P. Free energy calculations: An efficient adaptive biasing potential method. *J. Phys. Chem. B* **2010**, *114* (17), 5823–5830.

(47) Bartels, C.; Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.* **1997**, *18* (12), 1450–1462.

(48) Maragakis, P.; van der Vaart, A.; Karplus, M. Gaussian-mixture umbrella sampling. *J. Phys. Chem. B* **2009**, *113* (14), 4664–4673.

(49) Rosso, L.; Abrams, J. B.; Tuckerman, M. E. Mapping the backbone dihedral free-energy surfaces in small peptides in solution using adiabatic free-energy dynamics. *J. Phys. Chem. B* **2005**, *109* (9), 4162–4167.

(50) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12562–12566.

(51) Ensing, B.; De, V. M.; Liu, Z.; Moore, P.; Klein, M. L. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.* **2006**, *39* (2), 73–81.

(52) Vymetal, J.; Vondrasek, J. Metadynamics as a tool for mapping the conformational and free-energy space of peptides: The alanine dipeptide case study. *J. Phys. Chem. B* **2010**, *114* (16), 5632–5642.

(53) Seabra, G. d. M.; Walker, R. C.; Elstner, M.; Case, D. A.; Roitberg, A. E. Implementation of the SCC-DFTB method for hybrid QM/MM simulations within the amber molecular dynamics package. *J. Phy. Chem. A* **2007**, *111*, 5655–5664.

(54) Mu, Y. G.; Kosov, D. S.; Stock, G. Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. *J. Phys. Chem. B* **2003**, *107* (21), 5064–5073.

(55) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. Assessing the accuracy of metadynamics. *J. Phys. Chem. B* **2005**, *109* (14), 6714–21.

(56) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.* **1997**, *119* (25), 5908–5920.

(57) 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* (2), 785–9.

(58) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157* (3), 200–6.

(59) Becke, A. D. A new mixing of Hartree-Fock and local-density-functional theories. *J. Chem. Phys.* **1993**, *98* (2), 1372–7.

(60) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55* (1), 117–29.

(61) Miertus, S.; Tomasi, J. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chem. Phys.* **1982**, *65* (2), 239–45.

(62) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Ab initio* study of solvated molecules: a new implementation of the polarizable continuum model. *Chem. Phys. Lett.* **1996**, *255* (4,5,6), 327–335.

(63) 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.; and Pople, J. A. *Gaussian 03*, revision B.05; Gaussian, Inc.: Wallingford, CT, 2004.

(64) Wong, M. W. Vibrational frequency prediction using density functional theory. *Chem. Phys. Lett.* **1996**, *256* (4,5), 391–399.

(65) http://www.westcenter.usp.edu/code.

(66) Toukan, K.; Rahman, A. Molecular-dynamics study of atomic motions in water. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31* (5), 2643–8.

(67) Ahlborn, H.; Ji, X.; Space, B.; Moore, P. B. A combined instantaneous normal mode and time correlation function description of the infrared vibrational spectrum of ambient water. *J. Chem. Phys.* **1999**, *111* (23), 10622–10632.

(68) Siu, S. W. I.; Vacha, R.; Jungwirth, P.; Boeckmann, R. A. Biomolecular simulations of membranes: Physical properties from different force fields. *J. Chem. Phys.* **2008**, *128* (12), 125103–125115.

(69) Feig, M.; Pettitt, B. M. Experiment vs force fields: DNA conformation from molecular dynamics simulations. *J. Phys. Chem. B* **1997**, *101* (38), 7361–7363.

(70) Feller, S. E.; Pastor, R. W.; Rojnuckarin, A.; Bogusz, S.; Brooks, B. R. Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.* **1996**, *100* (42), 17011–17020.

(71) Norberg, J.; Nilsson, L. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.* **2000**, *79* (3), 1537–1553.

(72) Feig, M. Is Alanine dipeptide a good model for representing the torsional preferences of protein backbones. *J. Chem. Theory Comput.* **2008**, *4* (9), 1555–1564.

(73) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. DFT studies on helix formation in N-acetyl-(L-alanyl)n-N'-methylamide for n = 1−20. *Chem. Phys.* **2000**, *256* (1), 15–27.

(74) Jalkanen, K. J.; Elstner, M.; Suhai, S. Amino acids and small peptides as building blocks for proteins: comparative theoretical and spectroscopic studies. *THEOCHEM* **2004**, *675* (1−3), 61–77.

(75) Grenie, Y.; Avignon, M.; Garrigou-Lagrange, C. Molecular structure study of dipeptides isolated in an argon matrix by infrared spectroscopy. *J. Mol. Struct.* **1975**, *24* (2), 293–307.

(76) Ramachandran, G. N.; Sasisekharan, V. Conformation of polypeptides and proteins. *Adv. Protein Chem.* **1968**, *23*, 283–438.

CT100395N

JCTC Journal of Chemical Theory and Computation

Theoretical Investigations on the Conformation of the β-D-Arabinofuranoside Ring

Hashem A. Taha,[†] Pierre-Nicholas Roy,[‡] and Todd L. Lowary*,[†]

Department of Chemistry and Alberta Ingenuity Centre for Carbohydrate Science, Gunning-Lemieux Chemistry Centre, University of Alberta, Edmonton, AB, Canada T6G 2G2 and Department of Chemistry, University of Waterloo, Waterloo, ON, Canada N2L 3G1

Received August 12, 2010

Abstract: A method for the conformational analysis of furanose rings that involves the prediction of $^3J_{H,H}$ that can be compared directly to experimental values is investigated. This method, which differs from the traditional PSEUROT approach for conformational studies of furanose rings, was previously applied to a number of α-D-arabinofuranosides and enabled the direct comparison of $^3J_{H,H}$ values to those obtained from NMR spectroscopy. In this paper, the use of this approach to study the conformational preferences of oligosaccharides containing β-linked arabinofuranose residues is reported. Density functional theory (DFT) calculations were carried out to derive Karplus relationships that are specifically tailored for these ring systems. In addition, probability distributions obtained from GLYCAM/AMBER molecular dynamics simulations were employed to calculate $^3J_{H,H}$ values from these Karplus relationships. However, unlike the results obtained with α-arabinofuranosides, the $^3J_{H,H}$ values computed for β-arabinofuranosides agreed poorly with experimental values. This prompted the exploration of other methodologies including reevaluation and optimization of the initial MD protocol, use of various force field models, and recalculation of the DFT-derived coupling profiles using an optimized basis set. After extensive investigations, we established that the conformer distributions obtained from MD simulations with the GLYCAM force fields and the furanoside-specific CHARMM force field in combination with the DFT Karplus equations, determined using an augmented basis set (B3LYP/aug-cc-pVTZ-J), produced the best agreement compared to experimental $^3J_{H,H}$ values. Using these protocols, there is relatively good agreement in $^3J_{H,H}$ for all coupling pathways with the exception of $^3J_{2,3}$ and $^3J_{3,4}$, which are underestimated.

Introduction

Furanose (or five-membered ring) carbohydrates are important constituents of a number of glycoconjugates in many microorganisms.[1−4] Our group has a long-standing interest in conformational analysis of furanoside-containing polysaccharides found in the complex cell wall of the pathogenic species *Mycobacterium tuberculosis*, the causative agent of tuberculosis.[5−8] Due to the critical role that these glycoconjugates play in the viability and virulence of mycobacteria,[9] it is essential to study their conformation in order to understand their biological functions.

Furanose rings assume various twist (T) and envelope (E) conformations that can be depicted using the pseudorotational wheel (Figure 1). Each conformer is described by its Altona−Sundaralingam (AS) phase angle of pseudorotation (P), which represents the atoms that are displaced from the plane, and its AS puckering amplitude (ϕ_m), a measure of the maximum displacement from the planar ring form. Given five endocyclic torsion angles of a particular conformer, P and ϕ_m can be calculated.[10] These conformers interconvert readily because of the relatively low-energy barriers separat-

* Corresponding author. E-mail: tlowary@ualberta.ca.
[†] University of Alberta, Edmonton.
[‡] University of Waterloo.

10.1021/ct100450s
Published on Web 12/30/2010

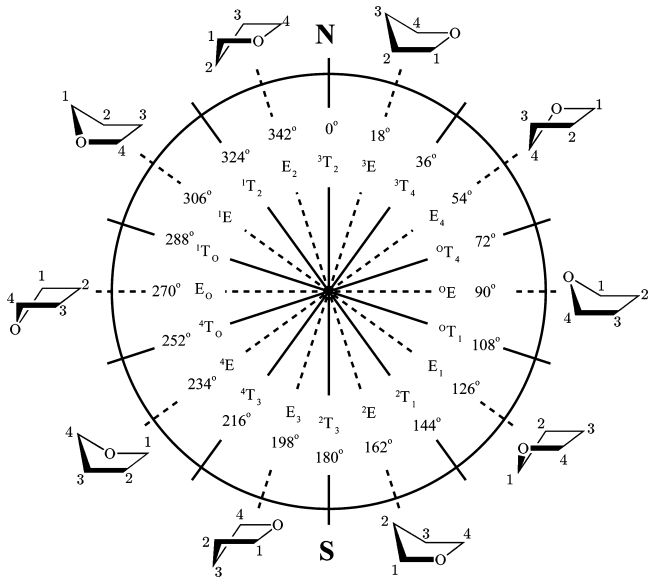


Figure 1. Pseudorotational itinerary for a D-furanose ring.

ing them (<5 kcal mol^{-1}).[11] This ring flexibility poses a challenge for the theoretical description of furanosides as both ring torsion angles as well as any exocyclic dihedral angles must be considered.

NMR spectroscopy has played a key role in the determination of the solution conformation of carbohydrates.[12−14] In particular, for furanosides, three-bond hydrogen−hydrogen coupling constants ($^3J_{H,H}$) obtained from NMR spectroscopy are commonly used in conjunction with a computer program, PSEUROT, to predict their conformational preferences.[10,15−18] This program assumes an equilibrium between two low-energy conformers, often located in the northern and southern hemispheres of the pseudorotational wheel, which interconvert through pseudorotation (Figure 1). This program takes experimental $^3J_{H,H}$ values for the ring hydrogens and calculates, using the appropriate Karplus relationships, two conformations and their mole fractions that fit the data the best. Although PSEUROT has been commonly employed for conformational analysis of five-membered rings, there are drawbacks to its use. For example, the two-state model is not valid in all cases, and the analysis may sometimes provide physically unrealistic conformations.[8,19,20] As an alternative, we have used theoretical models, such as molecular dynamics (MD) simulations together with density functional theory (DFT) calculations, to study conformation and dynamics.[6−8]

In a previous investigation,[8] we reported MD simulations of a number of oligosaccharides containing α-arabinofuranose (α-Ara*f*) residues. *β*-Arabinofuranose (*β*-Ara*f*) moieties are also found in nature, and these glycosidic residues play important roles within the cell wall structure of *M. tuberculosis*. In fact, *β*-Ara*f* residues (e.g., **1**−**5**, Figure 2) are usually found at the periphery of mycobacterial cell wall polysaccharides and are typically substituted with other groups that play key roles in the survival and pathogenicity of the organism.[1] In the arabinogalactan (AG), this group is esterified with mycolic acids,[1] while in the lipoarabinomannan (LAM), this position is capped by short mannopyranosyl oligosaccharides that are important in interactions with human mannose binding receptors.[21−23] One of our interests

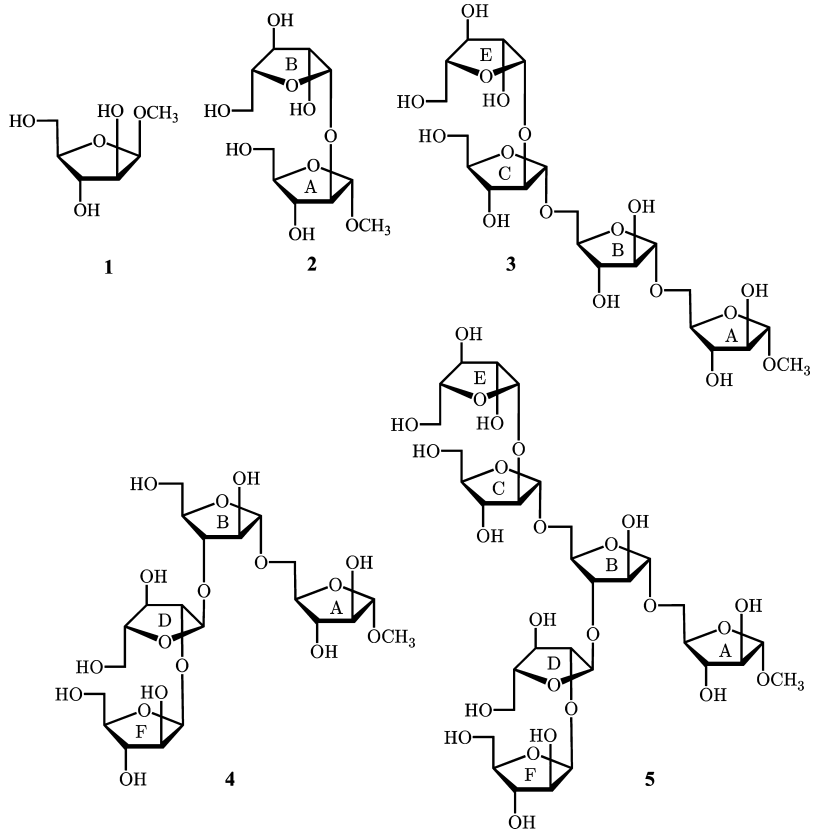


Figure 2. Studied *β*-Ara*f*-containing molecules.

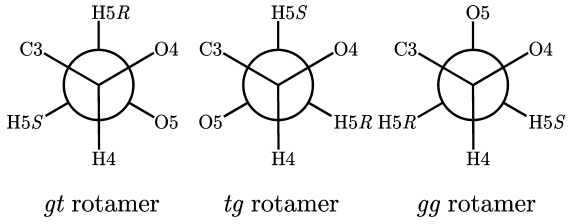


Figure 3. Definition of staggered rotamers about the C4−C5 bond.

is a hexasaccharide motif found at the nonreducing end of AG and LAM that is comprised of both α- and β-Ara*f* residues (**5**, Figure 2). It has been suggested that this hexasaccharide plays an important role in a number of immunological events that occur upon infection by mycobacteria.[24,25] For example, we have demonstrated that this motif is recognized by the anti-LAM antibody CS-35,[26,27] and thus this structure elicits an immune response.

Previously, we reported the use of the AMBER/GLYCAM approach to study the conformation of methyl β-D-arabinofuranoside (**1**, Figure 2), and in the course of these studies we demonstrated that the water model used had an important influence on the ability of this method to reproduce experimentally determined conformer populations.[7] More recently, we reported an alternative method to study the conformational preferences of α-Ara*f* systems.[8] This protocol involves the use of probability distributions from MD simulations to calculate Boltzmann-averaged $^3J_{H,H}$ values in combination with DFT-derived Karplus equations. The resulting coupling constants can be directly compared to those obtained from NMR spectroscopy. Better agreement with experiment was found using the DFT-derived Karplus equations compared to the use of the empirical Haasnoot−Altona Karplus relationship.[28] This approach provided an alternative to the use of PSEUROT for studying furanose conformation; notably, it does not require the two-state assumption.

In concert with MD simulations performed on oligosaccharides containing α-Ara*f* residues,[8] the use of the AMBER/GLYCAM approach is employed here to study the conformation of oligofuranosides containing β-Ara*f* residues (**1**−**5**, Figure 2). Although the use of this method was successful in probing the conformation of α-Ara*f* gycosides, its use in probing β-Ara*f* conformation proved problematic. In the present report, we investigate the potential sources of these problems, and a discussion of the various methods employed toward finding solutions is included.

Nomenclature

The three ideally staggered rotamers about the C4−C5 bond (gt, tg, and gg) in the Ara*f* residues are defined as shown in Figure 3.

Methods

DFT $^3J_{H,H}$ Coupling Profiles. In a manner similar to previously reported for methyl α-D-arabinofuranoside,[8] 10 envelope conformers of methyl β-D-arabinofuranoside (**1**) corresponding to all envelope structures indicated on the pseudorotational wheel (Figure 1) were constructed. For each envelope structure, three C4−C5 rotamers (gt, tg, gg) and three C5−O5 rotamers ($\psi = 180°$, $-60°$, and $60°$, where ψ is defined by the H5−O5−C5−C4 torsion angle) were generated, resulting in a total of 90 conformations. The geometries of all 90 conformations were then optimized with Gaussian 03[29] using the B3LYP functional[30] with the 6-31G* basis set. The torsion angle representing the four-atom plane of each envelope conformer was fixed at 0° to maintain the envelope structure. For example, the E_0 conformer was generated by fixing C1−C4 in the plane. All other geometric parameters were allowed to vary during the geometry optimizations.

DFT calculations of the spin−spin coupling constants in **1** were initially performed using Gaussian 03[29] at the B3LYP/cc-pVTZ level of theory.[30,31] All four contributions to the $^3J_{H,H}$ were computed (Fermi contact, diamagnetic spin orbit, paramagnetic spin orbit, and spin dipolar). The resulting *J* data were extracted for all conformations (see Table S-1 in the Supporting Information for complete coupling constant data).

The same spin−spin coupling calculations were also performed with an augmented basis set (aug-cc-pVTZ-J) that contains additional primitive s and p functions (compared to cc-pVTZ) and has been optimized for calculation of spin−spin coupling constants.[32−34] For comparison, a basis set [5s2p1d|3s1p] developed in the Serianni and Carmichael groups, designed to recover the Fermi contact contribution to the coupling,[35] was also employed. This basis set has been shown to provide good agreement with experimental $^3J_{H,H}$.[35] In addition, calculations with this basis set were much faster compared to the aug-cc-pVTZ-J calculations.

The Marquardt−Levenberg nonlinear least-squares algorithm[36] was used to fit the acquired coupling constants to the following truncated Fourier series in the H,H dihedral angle, ϕ:[37]

$$^3J_{H,H} = a + b\cos(\phi) + c\cos(2\phi) \quad (1)$$

The coefficients $a-c$ are obtained, corresponding to the five $^3J_{H,H}$ coupling pathways in **1**. In the particular cases of $^3J_{1,2}$, $^3J_{2,3}$, and $^3J_{4,5R}$, a phase shift to the dihedral angle (ϕ) was required to obtain improved fits.

GLYCAM04 MD Simulations. Initial simulations of **1** were carried out using the PMEMD implementation in the AMBER 10 suite of programs[38] with the AMBER force field and the GLYCAM carbohydrate parameter set (version 04f).[39] A four-step equilibration scheme was performed on **1** with an initial minimization step where the sugar was held fixed and the positions of water molecules were relaxed. A subsequent minimization allowed for all atoms to move for 50 steps of steepest descent followed by 950 steps of conjugate gradient to minimize the system as a whole.

Once sufficiently relaxed, the system underwent 100 ps of simulated annealing. The volume was kept constant, and the SHAKE[40] algorithm was used to constrain bonds involving hydrogen atoms. The final step before production dynamics involved equilibration of the physical parameters, such as temperature, pressure, and density of the system. This equilibration period was run over 240 ps using NPT

conditions, where temperature and pressure were held constant using a constant temperature thermostat with the weak coupling algorithm ($ntt = 1$)[41] and a constant pressure barostat with isotropic position scaling ($ntp = 1$), respectively.

The production phase was run for 250 ns under identical NPT conditions as the final equilibration step. This longer simulation time was chosen to ensure sufficient equilibration of the system and proper convergence. SHAKE was used, and long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) algorithm[42,43] with a cutoff of 8 Å. Coordinates were printed to the trajectory file every 1000 steps (every 2 ps).

GLYCAM06 MD Simulations. MD simulations of **1**−**5** were also performed using the GLYCAM06 force field[44] and the AMBER 10 suite of programs.[38] Oligosaccharides **2**−**5** were constructed from multiple units of the α and β anomers of **1** using additive atomic charges as described previously.[8] All other procedures are identical to the GLYCAM04 simulations.

Langevin Thermostat Simulations. MD simulations of **1** were carried out using a Langevin dynamics temperature regulation scheme[45] ($ntt = 3$) with a collision frequency (γ) of 2.0 ps^{-1} to address the validity of the algorithm used for maintaining a constant temperature throughout the simulations. Other simulation parameters remained unchanged.

Biased MD Simulations. A biased set of 200 conformations having P values in the range of $P = -5°$ to 30° was generated from a 50 ns MD simulation. Partial atomic charges for these conformers were calculated as previously reported;[6,7] the procedure and the resulting charges are included in the Supporting Information (Table S-2 and related discussion). Using this biased set of charges, MD simulations of **1** were carried out as before with no changes to the parameters indicated above.

CHARMM MD Simulations. MD simulations of **1** were also performed with the CHARMM program[46] in the constant pressure-constant temperature (NPT) ensemble using a Nosé−Hoover thermostat[47,48] with a reference temperature of 300 K and a Langevin piston barostat[49] with a reference pressure of 1 atm. The system was built using the force field parameters reported by Hatcher et al. for aldopentofuranosides[50] and was solvated via the CHARMMing web interface[51] with a cubic solvation of TIP3P water[52] molecules with a crystal dimension of 17.57 Å. The system then underwent 50 steps of steepest descent and 950 steps of conjugate gradient minimization, which was followed by a 100 ps period of gradual heating to a final temperature of 300 K. The system was equilibrated for 240 ps under NPT conditions. The production dynamics were run for 250 ns at 300 K, and the SHAKE algorithm was used to constrain all hydrogen atom bonds to their equilibrium length and to maintain rigid TIP3P water geometry. The long-range electrostatic interactions were treated with the PME summation.

QM/MM Simulations. Hybrid quantum mechanical and molecular mechanical (QM/MM) simulations of **1** were performed in a cubic box of 264 TIP3P water molecules[52] using the SANDER module in the AMBER 10 suite of programs.[38] The carbohydrate was treated using the semiempirical PM3CARB-1 parameter set,[53] and the solvent molecules were modeled classically. The PM3CARB-1 QM level of theory has been shown[53] to provide improved predictions for intramolecular hydrogen bonds, which are essential for correctly describing carbohydrates in an aqueous environment. Moreover, this level of theory was shown to improve predictions of structure and energetics of small carbohydrate analogues when compared to PM3.[53,54] In our simulations, there are no bonds that cross the QM/MM boundary, and therefore, hydrogen link atoms were not required (i.e., there are no covalent bonds between QM and MM atoms). Preparation of the system for production dynamics included a 1000 step minimization, followed by a 100 ps simulated annealing period and a 240 ps equilibration. All steps were run using QM/MM. The SHAKE algorithm was used to constrain all hydrogen atom bonds, and long-range electrostatics were treated with the PME algorithm using a cutoff of 8 Å. Coordinates were printed to the trajectory file every 500 steps (every 1 ps).

GROMACS MD Simulations. MD simulations of **1** were also performed using the GROMACS program[55] together with the GROMOS96 force field[56] and the SPC/E water model.[57] Partial atomic charges calculated using our modified GLYCAM approach (as described previously)[6,7] were used. The simulated system was composed of one molecule of **1** surrounded by 431 water molecules in a cubic box simulated under periodic boundary conditions. Newton's equations of motion were integrated using the GROMACS MD integrator with a 2 fs time step. The LINCS algorithm[58] was applied to constrain all bond lengths. The simulations were carried out in the NPT ensemble (at a constant temperature of 300 K and a pressure of 1 atm). The temperature and pressure were maintained constant using the weak-coupling Berendsen thermostat[41] and the Berendsen barostat via isotropic coordinate scaling.[41] The PME algorithm was used for treatment of electrostatics with a cutoff of 8 Å. Prior to production, the system was subjected to the three-step protocol (minimization, annealing, and equilibration) used in the other simulations. MD simulations were then conducted in the NPT ensemble for 200 ns, and data were collected every 2 ps.

$^3J_{H,H}$ from MD Conformer Ensembles. For an accurate comparison of the DFT/MD-derived $^3J_{H,H}$ values to experiment, ensemble averaging must be carried out. This was done by calculating $^3J_{H,H}$ values for each relevant fragment in compounds **1**−**5** using DFT-determined Karplus equations ($J(\phi)$, eq 2) in combination with the continuous probability distributions ($\rho(\phi)$, eq 2) of the respective $\phi_{H,H}$ obtained from MD simulations. In a similar manner as before,[8] these $^3J_{H,H}$ values were then ensemble-averaged using the following relation:

$$\langle J \rangle = \int_0^{360} J(\phi)\rho(\phi)\mathrm{d}\phi \qquad (2)$$

Direct DFT Coupling Calculations. A representative set of 200 conformations was extracted from a GLYCAM simulation of **1**. Coupling constants were then computed for each of these conformers in the Gaussian 03 program[29] using the B3LYP functional[30] and the [5s2p1d|3s1p] basis set.[35]

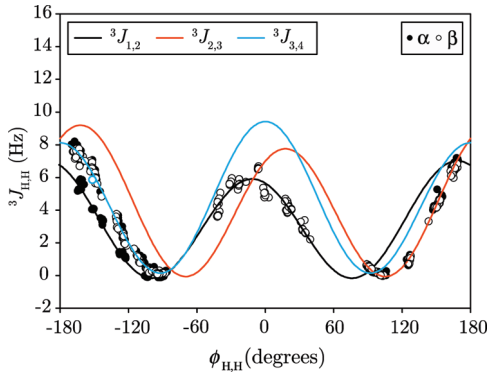


Figure 4. Karplus curves of $^3J_{1,2}$ (eq 3), $^3J_{2,3}$ (eq 4), and $^3J_{3,4}$ (eq 5) for methyl α-D-arabinofuranoside (filled circles, ●) and methyl β-D-arabinofuranoside (unfilled circles, ○) obtained from B3LYP/cc-pVTZ calculations.

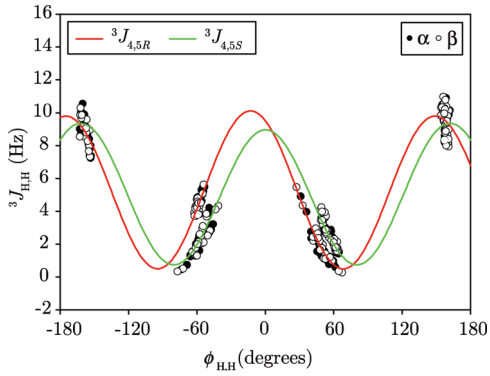


Figure 5. Karplus curves of $^3J_{4,5R}$ (eq 6) and $^3J_{4,5S}$ (eq 7) for methyl α-D-arabinofuranoside (filled circles, ●) and methyl β-D-arabinofuranoside (unfilled circles, ○).

The resulting $^3J_{H,H}$ values were then averaged over all 200 conformations.

Results and Discussion

Karplus Relationships for $^3J_{H,H}$ in D-Arabinofuranosides. In our previous report,[8] Karplus equations were developed for methyl α-D-arabinofuranoside, the α-anomer of **1**. In the interest of generalization to all D-arabinofuranosides, Karplus relationships that can be applied to both α- and β-Ara*f* residues were developed. To do this, $^3J_{H,H}$ values for both anomers of methyl D-arabinofuranoside were plotted together as a function of $\phi_{H,H}$. The data were fitted to obtain overall curves (Figures 4 and 5) and their corresponding parametrizations (eqs 3−7).

The ring protons display well-fitted curves (Figure 4 and eqs 3−5). The curve for $^3J_{3,4}$ is symmetrical about 0° and exhibits a global maximum of 10 Hz. In contrast, $^3J_{1,2}$ and $^3J_{2,3}$ curves are both shifted (nonsymmetry about 0°), and both required phase shifts for better fits (11° and −18°, respectively). In our previous report,[8] the $^3J_{1,2}$ Karplus curve for α-Ara*f* was not well parametrized around 0° because conformers with $\phi_{1,2}$ near 0° were not possible given the constraints of the ring system. With the addition of data points for **1**, this lack of parametrization at 0° has greatly improved.

$$^3J_{1,2}(\alpha,\beta) = 3.15 - 0.55\cos(\phi + 11°) + 3.30\cos(2\phi + 22°) \quad (R^2 = 0.98) \quad (3)$$

$$^3J_{2,3}(\alpha,\beta) = 4.21 - 0.72\cos(\phi - 18°) + 4.26\cos(2\phi - 36°) \quad (R^2 = 1.00) \quad (4)$$

$$^3J_{3,4}(\alpha,\beta) = 4.46 - 0.65\cos(\phi) + 4.31\cos(2\phi) \quad (R^2 = 0.99) \quad (5)$$

The exocyclic hydroxymethyl groups in both α- and β-Ara*f* exhibit similar coupling profiles and well-fitted Karplus curves (Figure 5, eqs 6−7). The curve for $^3J_{4,5R}$ is shifted from the $^3J_{4,5S}$ curve, and a phase shift of 15° was added to ϕ for a better fit.

$$^3J_{4,5R}(\alpha,\beta) = 5.22 - 0.15\cos(\phi + 15°) + 4.73\cos(2\phi + 30°) \quad (R^2 = 0.97) \quad (6)$$

$$^3J_{4,5S}(\alpha,\beta) = 4.94 - 0.20\cos(\phi) + 4.21\cos(2\phi) \quad (R^2 = 0.97) \quad (7)$$

MD/DFT-determined $3J_{H,H}$ in 1. Using the DFT-derived relationships determined above, we computed averaged $^3J_{H,H}$ values using the distribution of conformers that were obtained from MD simulations of **1** using the GLYCAM04 carbohydrate parameter set. Presented in Table 1 (G04) is a comparison of these calculated $^3J_{H,H}$ values for **1** with those measured by NMR spectroscopy. The $^3J_{H,H}$ values computed using the conformer ensemble obtained from simulations using the GLYCAM06 force field are also included in Table 1 (G06).

Analysis of these data reveals that the combination of the MD conformer ensembles and the DFT-derived equations is able to reproduce experimental $^3J_{1,2}$ and $^3J_{4,5S}$ values with near perfect accuracy. However, as was observed in the analysis of α-arabinofuranosides,[8] the computed $^3J_{4,5R}$ is underestimated compared with experiment (1.7 Hz deviation). This can be attributed to an underestimation of the gt rotamer population (i.e., the largest contributor to $^3J_{4,5R}$) by the MD simulations, which results in a lower overall average coupling constant.[8] For **2**−**5**, similar trends are observed for $^3J_{4,5}$ values (see Table S-4 in the Supporting Information) with deviations ranging from 1.7−2.6 Hz for $^3J_{4,5R}$ and 0−0.6 Hz for $^3J_{4,5S}$.

Unlike in the α-Ara*f* case,[8] the calculated $^3J_{2,3}$ and $^3J_{3,4}$ couplings in the present β-Ara*f* system also exhibit significant discrepancies compared to experimental values (deviations of 5.1 and 2.6 Hz, respectively). Analysis of the $^3J_{H,H}$ values obtained from the GLYCAM06 simulations reveals similar trends with slight improvements in the agreement of the ring couplings ($^3J_{1,2}$, $^3J_{2,3}$, and $^3J_{3,4}$).

To investigate the source of the discrepancies in $^3J_{2,3}$ and $^3J_{3,4}$ values, the respective dihedral angle distributions ($\phi_{2,3}$ and $\phi_{3,4}$) obtained from MD simulations of **1** were examined (Figure 6); for comparison, the distribution of $\phi_{1,2}$ is also included in Figure 6.

For $\phi_{1,2}$, two populations are observed in the MD conformer ensemble, with the most populated angles observed at −32° (67%) and 36° (33%). The DFT-derived

Table 1. $^3J_{H,H}$ Values (in Hz) in **1** Obtained from Experiment and from MD Simulation Conformer Populations[a]

	EXP	G04	G06	DDFT	B30	BIA	CHM	QMM	GRO
$^3J_{1,2}$	4.5	4.1	4.4	5.1	4.9	4.4	4.0	5.5	4.6
$^3J_{2,3}$	7.9	2.8	3.6	3.7	5.8	3.8	4.1	2.3	4.9
$^3J_{3,4}$	6.7	4.1	4.4	5.2	6.2	4.5	4.1	3.1	5.3
$^3J_{4,5R}$	6.7	5.1	5.0	5.7	4.5	4.9	5.0	6.4	4.8
$^3J_{4,5S}$	3.4	3.5	3.6	3.3	3.3	3.6	3.0	3.8	6.1

[a] EXP = experimental values; G04 = using GLYCAM04 conformer ensemble; G06 = using GLYCAM06 conformer ensemble; DDFT = average $^3J_{H,H}$ from 200 conformations from MD conformer ensemble; B30 = using a biased set of conformers having P values in the range of −30° to 30°; BIA = using conformer ensemble from simulations with biased set of atomic charges; CHM = using CHARMM conformer ensembles of **1**; QMM = using QM/MM conformer ensembles of **1**; and GRO = using GROMACS conformer ensembles of **1**.

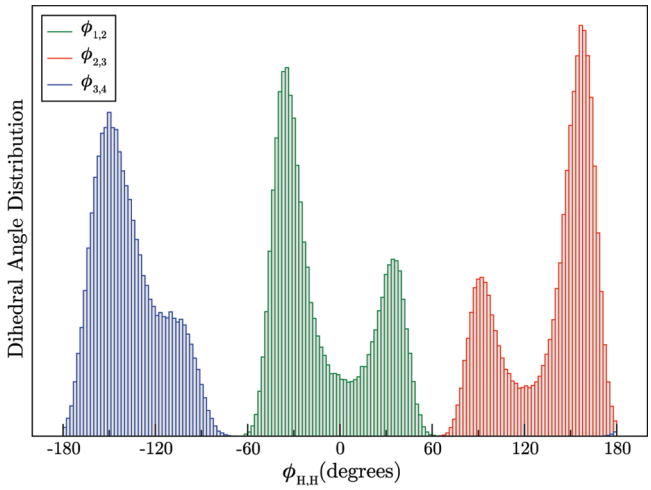


Figure 6. Distributions of ring protons obtained from the GLYCAM06 MD simulations of **1**.

Karplus curve obtained for $^3J_{1,2}$ along with the two-population distribution for $\phi_{1,2}$ produce coupling constants that is in good agreement with experiment. In contrast, the distributions for $\phi_{2,3}$ and $\phi_{3,4}$ each exhibit two-state populations that negatively impact the value of the average $^3J_{H,H}$.

The most populated state for $\phi_{2,3}$ (72%) is centered at 160° and that for $\phi_{3,4}$ is centered on −148° (77%). Using the DFT-determined Karplus curves for these coupling fragments, these dihedral angle distributions give a relatively large coupling constant because the respective hydrogen atoms are in a near-trans relationship. However, the second population of conformers for both angles is centered near 90° (or −90° for $\phi_{3,4}$). These distributions produce coupling values that are near 0 Hz. Therefore, ensemble averaging over the two populations in each case results in a low overall $^3J_{H,H}$ value. These two-state populations are more heightened in the GLYCAM04 distributions (see Figure S-1 in the Supporting Information), where essentially identical populations are observed for the $\phi_{2,3}$ angle (55:45) and similar populations for $\phi_{3,4}$ (66:34) compared to the GLYCAM06 distributions. This is reflected in the worse agreement between the calculated and experimental $^3J_{2,3}$ and $^3J_{3,4}$ values with the conformer ensemble obtained from the GLYCAM04 simulations (Table 1, G04).

It should be noted that the major conformation about each of these angles ($\phi_{1,2} = -32°$, $\phi_{2,3} = 160°$, and $\phi_{3,4} = -148°$) corresponds to conformers in the northern hemisphere of the pseudorotational wheel ($P = -31°-22°$). These structures are similar to the major conformer of **1**, as determined earlier using the PSEUROT approach, which predicts a conformational equilibrium biased heavily (~90:10) to a northern conformer ($E_2/^3T_2$; $P = -9$).[10] This structure is also in good agreement with the conformation of the ring in the crystal structure of the molecule.[59] Thus, the simulations appear to predict the correct major conformer but underestimates its population in the conformational equilibrium.

Direct DFT $^3J_{H,H}$ Calculation. To evaluate whether the errors in $^3J_{H,H}$ values stem from the Karplus curve fitting procedure, we investigated a direct method that bypasses this step. Rather than fitting $\phi_{H,H}$ and $^3J_{H,H}$ data computed for each fragment in **1** and using the generated equations with the MD conformer ensembles, a representative set of conformers was instead chosen, and $^3J_{H,H}$ values were directly calculated for this set. Subsequently, the final $^3J_{H,H}$ values were obtained by averaging over all conformers in the set. Table 1 (DDFT) shows the $^3J_{H,H}$ values obtained from these calculations and their comparison to $^3J_{H,H}$ values from our original approach as well as to experiment.

In terms of the $^3J_{4,5}$ couplings, this direct approach shows better agreement with experiment. A comparison of C4−C5 rotamer distributions reveals that the same relative trend is observed in both conformer ensembles (gt > gg > tg); however, small differences are detected (45 gt: 15 tg: 40 gg using the full conformer set vs 51 gt: 11 tg: 38 gg for the 200 conformer set). This finding suggests that, although not ideal, the conformer distribution could be refined (by selecting "better" conformations) to more accurately reproduce the experimental result. Analysis of the ring couplings ($^3J_{1,2}$, $^3J_{2,3}$, and $^3J_{3,4}$) demonstrated that the direct DFT approach results in better agreement with experiment for $^3J_{3,4}$ values but in poorer agreement for $^3J_{1,2}$; no differences in agreement were observed between the two sets for $^3J_{2,3}$. These results suggest that this small set of conformers is not sufficient to properly represent the distribution of H−H dihedral angles along each coupling pathway. Although a closer value was observed for $^3J_{3,4}$, this may be fortuitous from the random selection of conformers for this set.

It is obvious from these results that proper sampling of conformers was not achieved; certain conformers were not sampled sufficiently and others more than desired. It is probable that a larger conformer set, or a more biased selection of the set, is required to obtain better agreement. This is, however, a rather unsatisfying approach, which requires insight into the conformation of the molecule before it is studied.

Langevin Thermostat Simulations. Given the results presented above, we reevaluated our previously employed protocol[6−8] for carrying out the MD simulations of Ara*f* rings. In addition to the simulations performed using the Berendsen thermostat (used in all our previous simulations), MD simulations of **1** were also carried out using the Langevin thermostat. However, upon analysis of the resulting conformer ensembles, negligible differences were observed in

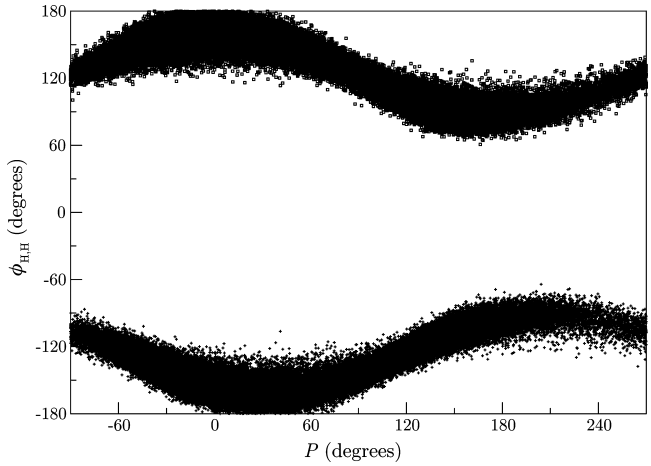


Figure 7. Plot of $\phi_{2,3}$ and $\phi_{3,4}$ as a function of P angle obtained from MD simulations of **1**.

both the rotamer populations as well as the distribution of ring conformations. For a discussion of the results, see Supporting Information (Figure S-2, Tables S-5 and S-6, and related discussion).

Biasing the MD Simulations. Further investigations to improve the predicted $^3J_{2,3}$ and $^3J_{3,4}$ values included the refinement of partial atomic charges to correspond to a biased set of conformations. In Figure 7, a plot of these couplings as a function of pseudorotational phase angle, P, is shown. In this plot, it can be seen that conformers that lie in the undesired $\phi_{2,3} = 90°$ or $\phi_{3,4} = -90°$ regions and give rise to $^3J_{H,H}$ near 0 Hz (Figure 6) correspond to P values in the southern hemisphere of the pseudorotational wheel (i.e., 90°−270°), and the desired conformers are in the northern P range of −90° to 90°.

To assess whether better agreement in $^3J_{2,3}$ and $^3J_{3,4}$ values can be obtained if only a particular set of conformers is used for the charge calculations, we extracted all conformations that adopted a P value falling in the desired range of −30° to 30° from the entire 250 ns trajectory. Indeed, we observed significant improvements in $^3J_{H,H}$ when using these conformations compared with the use of the entire trajectory (Table 1, compare B30 and G06).

With this result in hand, we modified our atomic charge calculation procedure so that ensemble averaging of charges was performed on only the conformations that adopted preferred P values. This was done in the hope that these biased charges would lead to an improved MD conformer ensemble and therefore better agreement in $^3J_{H,H}$. However, when the MD simulation of **1** was carried out using this biased set of charges, similar conformer ensembles were observed compared to the unbiased case; $^3J_{H,H}$ values were, therefore, also similar (Table 1, BIA). This result essentially reiterates the highly flexible nature of these furanoside systems. In fact, a time-dependence plot of the P angle (See Figure S-3 in Supporting Information) clearly indicates that even at short simulation times, all values of P can be readily visited (low-energy barriers); this is in contrast to the C4−C5 rotamers, which require long simulation times for proper sampling, especially the lower populated ones. These results are consistent with those previously described for MD simulations on arabinofuranosides[6−8,60] as well as reported DFT calculations on these systems, which have revealed that the barrier to pseudorotation is small (∼5 kcal/mol in the case of **1**)[61] and lower than the energy required to rotate about the C4−C5 bond.[62]

Table 2. MD Simulation Conformer Ensembles of **1** using the CHARMM Force Field

C4−C5 rotamers[a]			ring conformation[b]		
	current	Hatcher et al.[50]		current	Hatcher et al.[50]
X_{gt}	45	64	P_N	−7°	−15°
X_{tg}	15	6	%N	60	62
X_{gg}	40	30	P_S	160°	183°
			%S	40	38

[a] Experimental values: 57% gt, 8% tg, and 35% gg. [b] Experimental values: $P_N = -7°$, 86%; $P_S = 162°$, 14%.

In light of the above results, we questioned whether the use of the GLYCAM force fields exhibited some limitations for their application to furanosides. Therefore, we explored three alternate methods for carrying out simulations of **1**: MD simulations using a furanoside-specific force field[50] in the CHARMM program,[46] QM/MM simulations in the AMBER program,[63] and MD simulations using the GROMOS96 force field[56] in the GROMACS program.[64]

Use of the CHARMM Force Field. In a recent report,[50] Hatcher et al. reported an additive all-atom empirical force field parametrized for aldopentofuranoses and their methyl glycosides as well as for fructofuranose rings. Exocyclic rotamer populations and puckering distributions were predicted from aqueous-phase MD simulations of both anomers of methyl D-arabinofuranoside. Therefore, we utilized this force field for our own simulations of **1** and found that we qualitatively reproduced the results reported by Hatcher et al. (Table 2). The C4−C5 rotamer populations follow the same trend, though differing percentages are observed. This discrepancy is likely a result of the length of the MD simulations. In our protocol, simulation times of ≥200 ns were carried out to ensure convergence of C4−C5 rotamer populations,[6−8] whereas 20 ns simulations were employed in the Hatcher et al. report.[50] With regards to ring conformation, our simulations predicted north and south conformer populations that are almost identical to those previously reported[50] with the most populated conformers differing slightly.

With the MD conformer ensemble from the CHARMM simulations in hand, we carried out $^3J_{H,H}$ calculations as before, and the results are presented in Table 1 (CHM). This analysis reveals that, although slightly better agreement in $^3J_{2,3}$ is observed (4.1 Hz compared to 3.6 Hz), the remaining coupling constants exhibit similar or slightly worse agreement with experiment compared to those calculated with the GLYCAM06 conformer distributions. This discrepancy again can be attributed to the large percentage of southern conformers predicted by the simulations that correspond to near perpendicular $\phi_{2,3}$ and $\phi_{3,4}$ dihedral angles (see Figures S-3 and S-4 in the Supporting Information for a plot of the distributions). The near zero couplings resulting from these conformers give rise to a low overall $^3J_{H,H}$.

QM/MM Simulations of 1. The next method attempted toward a potential solution to the discrepancies in predicted

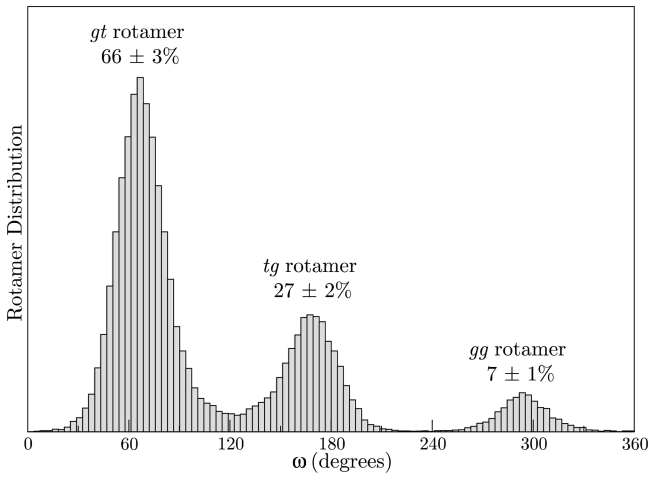


Figure 8. Histogram plot of the C4−C5 rotamer distributions obtained from QM/MM simulations of **1**.

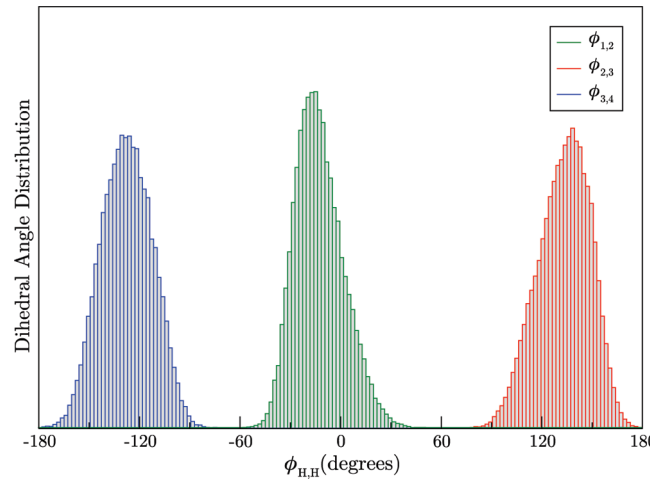


Figure 9. Distributions of ring protons obtained from the QM/MM simulations of **1**.

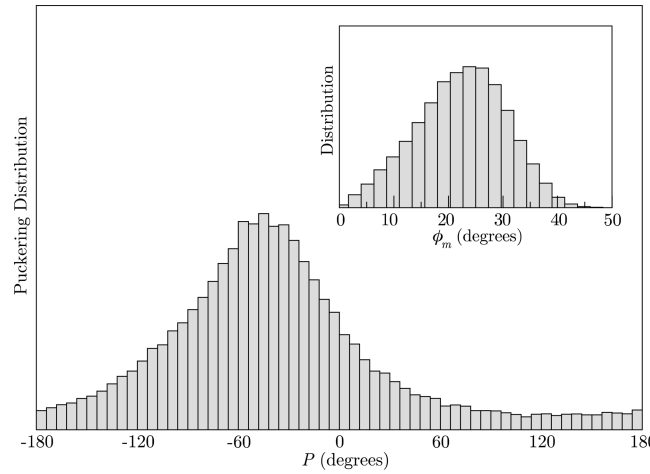


Figure 10. Distribution of the pseudorotational phase angle, *P*, and the puckering amplitude, ϕ_m (inset), obtained from QM/MM simulations of **1**.

$^3J_{H,H}$ values involved the use of a combined QM/MM approach where **1** was treated using QM, and the solvent was modeled with MM. The PM3CARB-1 parameter set was used in these simulations as it has shown improved prediction of intramolecular hydrogen-bond strength, ring conformation, and energetics compared to PM3.[53,54] In previous reports, this parameter set was used to accurately predict hydroxymethyl group conformation in gluco- and galactopyranose using QM/MM simulations[54] as well as adequate prediction of glycosidic linkage conformation in three disaccharides: (β-D-glucopyranosyl-(1→4)-β-D-glucopyranose, α-D-glucopyranosyl-(1→4)-α-D-glucopyranose, and α-D-galactopyranosyl-(1→4)-α-D-galactopyranose).[65]

Using the same convergence criteria as we employed previously (i.e., errors of ≤3% in C4−C5 rotamer populations), we observed that a simulation time of 100 ns was sufficient for proper convergence in these QM/MM simulations (see Figure S-6 in the Supporting Information for convergence plot). Figure 8 shows a histogram of the resulting rotamer distributions about the C4−C5 bond in **1**.

Integration of the peaks in the histogram produces a distribution of 66:27:7 for the gt:tg:gg rotamers. Unlike in previous MD simulations of **1**, this trend in the populations (gt > tg > gg) contradicts the experimental result (gt > gg > tg). The gg rotamer is found to be the least populated, indicating that the gauche effect,[66] the preference for adjacent electronegative substituents along a two-carbon fragment to adopt the gauche orientation, is not properly considered in the calculations. Moreover, hydrogen-bond analysis of the resulting conformer ensemble (see Table S-7 in the Supporting Information) showed no significant occupancy of intramolecular hydrogen bonds.

Upon analysis of the $^3J_{H,H}$ using the B3LYP/cc-pVTZ-determined Karplus equations with the conformer ensemble from the QM/MM simulations, we observed excellent agreement in the $^3J_{4,5}$ couplings (Table 1, QMM). This result suggests errors in the experimental model used to calculate rotamer populations. Having correctly reproduced the $^3J_{4,5}$ obtained from NMR spectroscopy, it would appear that these QM/MM simulations are the ideal choice for determining the hydroxymethyl group conformation in this system. However, as will be seen later when more accurate Karplus equations are used, this is not the case.

Analysis of the ring $^3J_{H,H}$ (Table 1, QMM) revealed much worse agreement with experiment than those calculated from the other MD simulations. To understand this discrepancy, we again looked at the dihedral angle distributions of the ring protons obtained from these QM/MM simulations (Figure 9). For all ring protons, the $H_x-C_x-C_{x+1}-H_{x+1}$ distributions each exhibit single-state populations, which is in contrast to what was observed in the classic MD simulations (See Figure 6).

For $\phi_{1,2}$, the distribution is centered about −13°, which corresponds to a larger $^3J_{1,2}$ than observed in experiment (see Karplus curve for $^3J_{1,2}$, above). Similarly, the distributions of $\phi_{2,3}$ and $\phi_{3,4}$ show the most probable dihedral angles at 140° and −127°, respectively, both of which correspond to low $^3J_{H,H}$ values (2.7 and 2.9 Hz, respectively). Upon examination of the distributions of ring conformations (Figure 10), the source of these discrepancies in the dihedral angle distributions became clear.

The distribution of *P* produces an average of −37° that corresponds to a conformation in the northwestern region

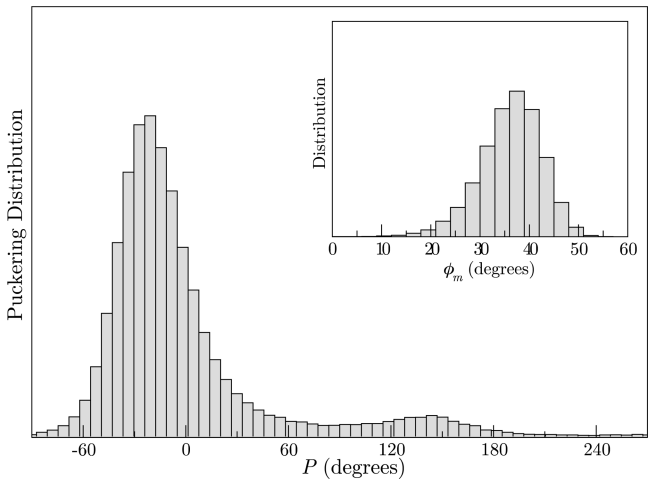


Figure 11. Distribution of the pseudorotational phase angle, *P*, and the puckering amplitude, ϕ_m (inset), obtained from GROMACS simulations of **1**.

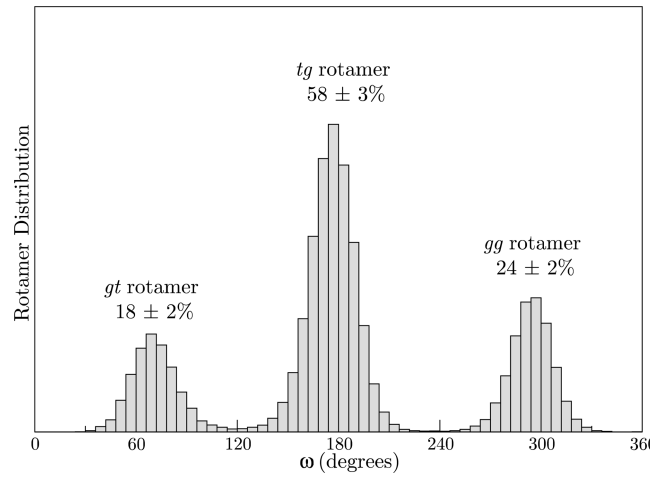


Figure 12. Histogram plot of the C4−C5 rotamer distributions obtained from GROMACS simulations of **1**.

of the pseudorotational wheel ($^1E/^1T_2$). This value is comparable to the northern *P* values predicted from the other simulations. More interestingly, however, is the predicted puckering amplitude, ϕ_m, which gives an average of 22°. A statistical analysis of a large number of β-D-furanoside X-ray structures suggest the β-Ara*f* ring adopts an optimal ϕ_m of 38°.[59,67] Therefore the QM/MM simulations predict a ring that is too flat, which we propose results in undesired ring H−C−C−H dihedral angle distributions. From these results, we can conclude that although the use of PM3CARB-1 in QM/MM simulations of **1** appears to correctly predict hydroxymethyl group conformation, it is not sufficient for determining ring conformation in **1**.

GROMACS Simulations of 1. In a recent report, unconstrained MD simulations of 2-*O*-sulfo-α-L-iduronic acid (IdoA2S) were carried using the GLYCAM06 and GROMOS96 force fields to investigate their ability to reproduce conformational distributions of the idopyranose ring, another flexible monosaccharide.[68−71] It was found that the predicted ring conformation using GROMOS96 was in better agreement with experiment than the use of the GLYCAM06 force field. Moreover, the predicted hydroxymethyl group conformation was similar in both cases. Therefore, to probe its performance in our systems, the GROMOS96 force field was utilized in MD simulations of **1**.

The resulting distribution of ring conformations as well as the C4−C5 rotamer populations is presented in Figures 11 and 12, respectively. Analysis of the puckering (Figure 11) shows a heavily biased distribution (92%) of northern conformers centered about $P_N = -14°$ and a small population (8%) of southern conformers centered on $P_S = 144°$. Moreover, the predicted puckering amplitude agrees well with previous simulations (with the exception of QM/MM), DFT theory calculations,[11] and X-ray data.[59]

Analysis of the ring $^3J_{H,H}$ values (Table 1, GRO) shows that with this puckering distribution, we indeed observe better agreement with experiment compared to the other simulations. Although the $^3J_{2,3}$ and $^3J_{3,4}$ values remain too small, the resulting values are now much closer to experiment compared to the other simulations, and we observe near perfect agreement in $^3J_{1,2}$.

Upon analysis of $^3J_{4,5}$, however, we observe significant deviation from experiment; the predicted values are now in reverse order. This result can be explained in a similar manner to that reported previously.[8] Looking at the C4−C5 rotamer distributions obtained from the MD simulations (Figure 12), it can be seen that the tg rotamer is predicted to be the most populated (58%). This rotamer is the largest contributor to $^3J_{4,5S}$, as opposed to the gt rotamer which produces the largest $^3J_{4,5R}$. Hence, the observed trend in $^3J_{4,5}$ couplings is a result of significant overestimation of the tg rotamer by the MD simulations.

More Accurate DFT Coupling Profiles. The significantly lower magnitudes in $^3J_{2,3}$ and $^3J_{3,4}$ compared to experiment prompted reinvestigation of the Karplus relationships calculated for the β-Ara*f* system. Upon analysis of all coupling constants calculated using the B3LYP/cc-pVTZ level of theory, we discovered that the maximum $^3J_{2,3}$ value that could be obtained was 6.8 Hz. This indicated that, regardless of which conformer ensemble is chosen, the predicted $^3J_{2,3}$ will not likely reach the experimental value of 7.9 Hz. Therefore, additional coupling profiles for **1** were computed using DFT calculations and an augmented basis set (aug-cc-pVTZ-J). This basis set has been optimized for calculation of spin−spin coupling constants and has been shown to produce accurate one-, two-, and three-bond *J* values in a number of small molecules containing electronegative substituents.[34] The spin−spin coupling constant data calculated using this augmented basis set were plotted as a function of the respective H−C−C−H torsion angles and fitted to eq 1 (the resulting curves along with their parametrizations are presented in the Supporting Information, Figure S-12 and equations S-1−5).

With these revised Karplus relationships in hand, we carried out calculations of averaged $^3J_{H,H}$ values using the distribution of conformers from all previous simulations of **1**. Analysis of the resulting data (Table 3, A) clearly shows that, overall, the $^3J_{H,H}$ magnitudes are larger than those obtained using the original DFT-derived Karplus relationships (eqs 3−7). In general, there is good agreement with

Table 3. $^3J_{H,H}$ Values (in Hz) for **1** using DFT-Derived Karplus Curves with Various MD Conformer Ensembles[a]

	EXP	G04	G06	CHM	QMM	GRO
A	B3LYP/aug-cc-pVTZ-J-Derived Karplus Curves					
$^3J_{1,2}$	4.5	4.9	5.1	4.7	6.7	5.4
$^3J_{2,3}$	7.9	3.5	4.6	5.3	2.9	6.3
$^3J_{3,4}$	6.7	5.2	5.5	5.1	3.8	6.6
$^3J_{4,5R}$	6.7	6.3	6.1	6.2	7.9	5.6
$^3J_{4,5S}$	3.4	4.0	4.1	3.3	4.5	7.4
B	B3LYP/5s2p1d|3s1p-Derived Karplus Curves					
$^3J_{1,2}$	4.5	4.6	4.9	4.5	6.2	5.1
$^3J_{2,3}$	7.9	3.2	4.3	4.8	2.6	5.9
$^3J_{3,4}$	6.7	4.8	5.1	4.8	3.4	6.2
$^3J_{4,5R}$	6.7	5.8	5.7	5.8	7.4	5.3
$^3J_{4,5S}$	3.4	3.7	3.8	3.2	4.2	7.0

[a] EXP = experimental values; G04 = using GLYCAM04 conformer ensembles of **1**; G06 = using GLYCAM06 conformer ensembles of **1**; CHM = using CHARMM conformer ensembles of **1**; QMM = using QM/MM conformer ensembles of **1**; and GRO = using GROMACS conformer ensembles of **1**.

experiment for $^3J_{1,2}$ (with the exception of QM/MM) and $^3J_{4,5S}$ (with the exception of GROMACS). Moreover, both of the GLYCAM and the CHARMM simulations predict comparatively accurate $^3J_{4,5R}$ values.

Overall, there is generally better agreement in $^3J_{2,3}$ and $^3J_{3,4}$ values; however, significant underestimation of the magnitudes remains. The GROMACS simulations predict the closest $^3J_{2,3}$ and $^3J_{3,4}$ values, although the trend remains in reverse compared to experiment, and $^3J_{1,2}$ has deviated away from the experimental value. Moreover, as discussed above, the GROMACS-predicted $^3J_{4,5}$ couplings are inconsistent with experiment.

The QM/MM simulations provide the worst agreement overall; using these new Karplus relationships, not a single $^3J_{H,H}$ shows reasonable agreement. In contrast, the best agreement in $^3J_{1,2}$, $^3J_{4,5R}$, and $^3J_{4,5S}$, and the correct trend in $^3J_{2,3}$ and $^3J_{3,4}$ is provided by the CHARMM simulations. For the sake of completeness, we also used Karplus equations using the Serianni−Carmichael [5s2p1d|3s1p] basis set to calculate $^3J_{H,H}$. In fact, similar results compared to the augmented basis set were obtained using this basis set (Table 3, B).

Conclusions

We report here the combined use of conformer ensembles obtained from MD simulations and DFT-derived Karplus relationships for subsequent calculation of $^3J_{H,H}$ ($^3J_{1,2}$, $^3J_{2,3}$, $^3J_{3,4}$, $^3J_{4,5R}$, and $^3J_{4,5S}$) as a conformational probe. This approach allows for the direct comparison of vicinal coupling constants obtained from NMR spectroscopy, thereby avoiding possible sources of errors encountered in the models used to analyze NMR data (e.g., the two-state model inherent in PSEUROT[15] or the "discrete" model).[72]

The coupling constant values calculated from the DFT-derived $^3J_{H,H}$ relationships for α-Ara*f* residues, as reported previously,[8] showed reasonable agreement with experiment. This result reiterates the ability of the AMBER/GLYCAM simulations to provide accurate conformer distributions of oligosaccharides containing α-Ara*f* rings. However, studies on the *β*-Ara*f* system using this approach displayed a number of difficulties.

Conformer ensembles obtained from MD simulations using the GLYCAM04 parameter set and the GLYCAM06 force field were used to calculate $^3J_{H,H}$ in **1**. The results show that reasonable agreement can be obtained for $^3J_{1,2}$ and $^3J_{4,5}$, but significant deviations are observed for $^3J_{2,3}$ and $^3J_{3,4}$. To understand this discrepancy, we evaluated the dihedral angle distributions predicted by the MD simulations along these fragments and found that a large population of conformers adopt near perpendicular $\phi_{2,3}$ and $\phi_{3,4}$ angles, which result in negligible $^3J_{H,H}$ values. These distributions arise from ring conformations that are present in the southern region of the pseudorotational wheel, which, on the basis of previous experimental work,[10,61,73] appear to be populated only to a small degree in solution. In fact, an analysis of $^3J_{H,H}$ for northern conformers showed significant improvements for the ring couplings over the use of the entire trajectory. However, when a set of northern-biased partial atomic charges was used in MD simulations of **1**, no change in the distribution of puckering was observed.

To find a potential solution to the discrepancy in $^3J_{H,H}$, we explored a direct DFT method, which avoids generating Karplus equations and instead calculates $^3J_{H,H}$ from a representative set of conformations. We envisioned that this protocol would circumvent any errors that may be introduced in the fitting procedure. The resulting $^3J_{H,H}$ from this method showed slightly better agreement with experiment, in general. A larger set of conformers may be required to accurately represent the phase space of this molecule. However, as the numbers of conformers required for good agreement with experiment increases, the practicality of this approach decreases due to the large cost of the DFT spin−spin coupling calculations. Potentially, the MD conformer ensemble can be tailored so as to reproduce the experimental data. This is, in principle, similar to a time-averaged restrained molecular dynamics (tar-MD) simulation where NMR restraints are used to bias the simulation to reproduce experimental data. This procedure was, in fact, recently used to study the conformation of a number of ribofuranose-based molecules.[19] However, this requires prior knowledge of NMR data and therefore would be insufficient for large oligosaccharides where experimental data can be difficult to obtain, due to spectral overlap. Furthermore, use of these approaches may hinder the development of an unbiased and general model to accurately probe the conformational preferences of these Ara*f* systems.

Different force fields were also investigated in simulations of **1** for their ability to predict accurate conformer ensembles. The recently developed CHARMM force field for aldopentofuranosides[50] predicted average $^3J_{H,H}$ values that were in similar agreement with experiment compared to GLYCAM06. The predicted C4−C5 rotamer populations as well as ring conformer distributions were similar to those reported by Hatcher et al.[50] Use of the GROMOS96 force field showed $^3J_{2,3}$ and $^3J_{3,4}$ values that are closer to experiment than those predicted by CHARMM, but the couplings along the C4−C5 fragment deviated significantly. The

predicted conformation of the hydroxymethyl group showed a distribution of 18:58:24 for gt:tg:gg. This result is peculiar, as the tg rotamer (which is the most populated in this case) lacks any stabilizing stereoelectronic effects that are present in the gt or gg rotamers, such as the gauche effect or intramolecular hydrogen bonds. Recent investigations on IdoA2S conformation showed that both GROMOS96 and GLYCAM06 force field are able to accurately predict hydroxymethyl group conformation.[74] In our simulations, however, these two force fields predict significantly different conformer distributions.

QM/MM simulations of **1** showed conflicting results. With the original DFT Karplus equations (from the B3LYP/cc-pVTZ calculations), these simulations appeared to have correctly predicted the hydroxymethyl group conformation. However, with the use of the augmented basis set or the Serianni−Carmichael basis set, both of which gave reasonable agreements using the classical MD conformer ensembles, the semiempirical QM/MM simulations resulted in contradictory results. It is possible that a different QM theory (other than PM3CARB-1) may be needed for better agreement. Alternatively, a DFT-MD approach that was recently applied to study the conformation of glucopyranose and all its epimers[75] may also be useful in this system. However, the use of this methodology for larger systems (such as **2**−**5**) is not practical from a computational perspective.

A final attempt at obtaining accurate $^{3}J_{H,H}$ from MD conformer ensembles involved the reevaluation of the DFT-derived Karplus equations. In the earlier work on α-Ara*f*-containing molecules, the coupling profiles generated using the Dunning cc-pVTZ basis set proved to be sufficient for predicting $^{3}J_{H,H}$.[8] However, for the present β-Ara*f* case, an augmented basis set (aug-cc-pVTZ-J) was required to obtain closer $^{3}J_{H,H}$ values; similar agreement was also observed using the Serianni−Carmichael [5s2p1d|3s1p] basis set. In comparison, use of the Serianni−Carmichael basis set offers a more superior method in terms of its relatively smaller size, and therefore, its more efficient calculation of spin−spin coupling profiles. In both cases, the CHARMM simulations appear to provide the best agreement in $^{3}J_{H,H}$ with experiment, although significant deviations were still observed, and GLYCAM06 performs similarly.

In conclusion, the range of simulation methods used here to model the β-Ara*f* ring demonstrated that the conformer populations obtained predict $^{1}H-^{1}H$ vicinal coupling constants that were in less good agreement with experiment, compared to previous investigations of α-Ara*f* rings. It is, of course, possible that other fixed charged force field models not attempted here, such as MM4,[76,77] could result in better agreement with experiment. Moreover, reparameterization of force field torsional functions may also be performed to obtain possibly better conformer distributions and work in this direction is currently underway. Other possibilities include the use of polarizable force field models to capture electronic polarization effects. The use of such force fields has not been required in the case of oligosaccharides containing pyranose residues.[78−82] However, as shown in previous reports,[6,7] fluctuations in fixed partial atomic charges in five-membered rings can change significantly as a function of ring pucker of both α- and β-Ara*f* rings, and therefore, inclusion of polarization in the model may result in better prediction of ring conformation. Models that include explicit treatment of electronic polarizability have been developed that can be used to treat alcohols.[83−88] Polarizable empirical force fields that are based on the classical Drude model[86,88] have also been reported for primary and secondary alcohols[89] as well as for linear and cyclic ethers.[90] Moreover, a general purpose polarizable model, AMOEBA, which replaces the fixed partial charge model with polarizable atomic multipoles through the quadrupole moments, has also been recently developed.[91] Therefore, we anticipate that a more accurate electrostatic representation together with force field torsional reparameterization would benefit the depiction of the conformational preferences of both α- and β-Ara*f* systems.

Acknowledgment. This work was supported by the Alberta Ingenuity Centre for Carbohydrate Science and the Natural Sciences and Engineering Research Council of Canada. H.A.T. thanks the Province of Alberta for a Queen Elizabeth II scholarship.

Supporting Information Available: Coupling constant data for Karplus curve fitting, biased partial atomic charges for **1**, $^{3}J_{H,H}$ values in **2**−**5**, couplings for different thermostats, time dependence of *P*, conformer distributions from CHARMM and QM/MM simulations, H-bond analysis from QM/MM simulations, and more accurate Karplus equations. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Brennan, P. J.; Nikaido, H. *Annu. Rev. Biochem.* **1995**, *64*, 29.

(2) de Lederkremer, R. M.; Colli, W. *Glycobiology* **1995**, *5*, 547.

(3) Peltier, P.; Euzen, R.; Daniellou, R.; Nugier-Chauvin, C.; Ferrières, V. *Carbohydr. Res.* **2008**, *343*, 1897.

(4) Richards, M. R.; Lowary, T. L. *ChemBioChem* **2009**, *10*, 1920.

(5) Lowary, T. L. *Curr. Opin. Chem. Biol.* **2003**, *7*, 749.

(6) Seo, M.; Castillo, N.; Ganzynkowicz, R.; Daniels, C. R.; Woods, R. J.; Lowary, T. L.; Roy, P.-N. *J. Chem. Theory Comput.* **2008**, *4*, 184.

(7) Taha, H. A.; Castillo, N.; Roy, P.; Lowary, T. L. *J. Chem. Theory Comput.* **2009**, *5*, 430.

(8) Taha, H. A.; Castillo, N.; Sears, D. N.; Wasylishen, R. E.; Lowary, T. L.; Roy, P. *J. Chem. Theory Comput.* **2010**, *6*, 212.

(9) Brennan, P. J. *Tuberculosis* **2003**, *83*, 91.

(10) Houseknecht, J.; Altona, C.; Hadad, C. M.; Lowary, T. L. *J. Org. Chem.* **2002**, *67*, 4647.

(11) Houseknecht, J.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2003**, *107*, 5763.

(12) Homans, S. W. Conformational Analysis in Solution by NMR. In *Carbohydrates in Chemistry and Biology*; Ernst, B., Hart, G. W., Sinay, P., Eds.; Wiley-VCH: New York, 2000; pp 947.

(13) Jimenez-Barbero, J.; Diaz, M. D.; Nieto, P. M. *Anti-Cancer Agents Med. Chem.* **2008**, *8*, 52.

(14) Kato, K.; Sasakawa, H.; Kamiya, Y.; Utsumi, M.; Nakano, M.; Takahashi, N.; Yamaguchi, Y. *Biochim. Biophys. Acta* **2008**, *1780*, 619.

(15) de Leeuw, F. A. A. M.; Altona, C. *J. Comput. Chem.* **1983**, *4*, 428.

(16) Hendrickx, P. M. S.; Martins, J. C. *Chem. Cent. J.* **2008**, *2*, 20.

(17) Thibaudeau, C.; Kumar, A.; Bekiroglu, S.; Matsuda, A.; Marquez, V. E.; Chattopadhyaya, J. *J. Org. Chem.* **1998**, *63*, 5447.

(18) Barchi, J. J.; Karki, R. G.; Nicklaus, M. C.; Siddiqui, M. A.; George, C.; Mikhailopulo, I. A.; Marquez, V. E. *J. Am. Chem. Soc.* **2008**, *130*, 9048.

(19) Hendrickx, P.; Corzana, F.; Depraetere, S.; Tourwe, D.; Augustyns, K.; Martins, J. *J. Comput. Chem.* **2010**, *31*, 561.

(20) Plavec, J.; Koole, L. H.; Chattopadhyaya, J. *J. Biochem. Biophys. Methods* **1992**, *25*, 253.

(21) Schlesinger, L. S. *Curr. Top. Microbiol. Immunol.* **1996**, *215*, 71.

(22) Nigou, J.; Gilleron, M.; Puzo, G. *Biochimie* **2003**, *85*, 153.

(23) Chatterjee, D.; Lowell, K.; Rivoire, B.; McNeil, M. R.; Brennan, P. J. *J. Biol. Chem.* **1992**, *267*, 6234.

(24) Chatterjee, D.; Roberts, A. D.; Lowell, K.; Brennan, P. J.; Orme, I. M. *Infect. Immun.* **1992**, *60*, 1249.

(25) Chatterjee, D.; Bozic, C. M.; McNeil, M.; Brennan, P. J. *J. Biol. Chem.* **1991**, *266*, 9652.

(26) Rademacher, C.; Shoemaker, G.; Kim, H.; Zheng, R.; Taha, H. A.; Liu, C.; Nacario, R.; Schriemer, D.; Klassen, J. S.; Peters, T.; Lowary, T. L. *J. Am. Chem. Soc.* **2007**, *129*, 10489.

(27) Murase, T.; Zheng, R. B.; Joe, M.; Bai, Y.; Marcus, S. L.; Lowary, T. L.; Ng, K. K. S. *J. Mol. Biol.* **2009**, *392*, 381.

(28) Altona, C.; Francke, R.; de Haan, R.; Ippel, J. H.; Daalmans, G. J.; Hoekzema, A. J. A. W.; van Wijk, J. *Magn. Reson. Chem.* **1994**, *32*, 670.

(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.; and Pople, J. A. *Gaussian 03*, revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.

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

(31) Helgaker, T.; Watson, M.; Handy, N. *J. Chem. Phys.* **2000**, *113*, 9402.

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

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

(34) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324.

(35) Stenutz, R.; Carmichael, I.; Widmalm, G.; Serianni, A. S. *J. Org. Chem.* **2002**, *67*, 949.

(36) Marquardt, D. W. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431.

(37) Haasnoot, C.; de Leeuw, F.; Altona, C. *Tetrahedron* **1980**, *36*, 2783.

(38) Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang; B. Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong; K. F. Paesani, F.; Vanicek, J.; Wu; X. Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*, University of California: San Francisco, 2008.

(39) Woods, R. J.; Dwek, R.; Edge, C.; Fraser-Reid, B. *J. Phys. Chem.* **1995**, *99*, 3832.

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

(41) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

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

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

(44) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622.

(45) Adelman, S. A.; Doll, J. D. *J. Chem. Phys.* **1976**, *64*, 2375.

(46) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.

(47) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(48) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(49) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613.

(50) Hatcher, E.; Guvench, O.; MacKerell, A. D. *J. Phys. Chem. B* **2009**, *113*, 12466.

(51) Miller, B. T.; Singh, R. P.; Klauda, J. B.; Hodošček, M.; Brooks, B. R.; Woodcock, H. L. *J. Chem. Inf. Model.* **2008**, *48*, 1920.

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

(53) McNamara, J. P.; Muslim, A. M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A. *Chem. Phys. Lett.* **2004**, *394*, 429.

(54) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2008**, *112*, 15450.

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

(56) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulations: the GROMOS96 Manual and User Guide*; Verlag der Fachvereine Hochschulverlag AG an der ETH Zürich: Zürich, Switzerland, 1996; pp 1.

(57) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.

(58) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463.

(59) Evdokimov, A.; Gilboa, A. J.; Koetzle, T. F.; Klooster, W. T.; Schultz, A. J.; Mason, S. A.; Albinati, A.; Frolow, F. *Acta Crystallogr., Sect. B: Struct. Sci.* **2001**, *57*, 213.

(60) Cros, S.; Hervé du Penhoat, C.; Pérez, S.; Imberty, A. *Carbohydr. Res.* **1993**, *248*, 81.

(61) Gordon, M.; Lowary, T. L.; Hadad, C. M. *J. Org. Chem.* **2000**, *65*, 4954.

(62) McCarren, P. R.; Gordon, M. T.; Lowary, T. L.; Hadad, C. M. *J. Phys. Chem. A* **2001**, *105*, 5911.

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

(64) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306.

(65) Stortz, C. A.; Johnson, G. P.; French, A. D.; Csonka, G. I. *Carbohydr. Res.* **2009**, *344*, 2217.

(66) Wolfe, S. *Acc. Chem. Res.* **1972**, *5*, 102.

(67) de Leeuw, H. P. M.; Haasnoot, C.; Altona, C. *Isr. J. Chem.* **1980**, *20*, 108.

(68) Angyal, S. J. *Aust. J. Chem.* **1968**, *21*, 2737.

(69) Angyal, S. J. *Angew. Chem., Int. Ed.* **1969**, *8*, 157.

(70) Angyal, S. J.; Kondo, Y. *Carbohydr. Res.* **1980**, *81*, 35.

(71) Angyal, S. J.; Pickles, V. A. *Aust. J. Chem.* **1972**, *25*, 1695.

(72) Džakula, Z.; Westler, W. M.; Edison, A. S.; Markley, J. L. *J. Am. Chem. Soc.* **1992**, *114*, 6195.

(73) Houseknecht, J.; Lowary, T. L. *J. Org. Chem.* **2002**, *67*, 4150.

(74) Gandhi, N. S.; Mancera, R. L. *Carbohydr. Res.* **2010**, *345*, 689.

(75) Schnupf, U.; Willett, J.; Momany, F. *Carbohydr. Res.* **2010**, *345*, 503.

(76) Allinger, N. L.; Chen, K.; Lii, J.; Durkin, K. A. *J. Comput. Chem.* **2003**, *24*, 1447.

(77) Lii, J.; Allinger, N. L. *J. Phys. Chem. A* **2008**, *112*, 11903.

(78) Brisson, J. R.; Uhrinova, S.; Woods, R. J.; van der Zwan, M.; Jarrell, H. C.; Paoletti, L. C.; Kasper, D. L.; Jennings, H. J. *Biochemistry* **1997**, *36*, 3278.

(79) Corzana, F.; Motawia, M. S.; Hervé du Penhoat, C.; Pérez, S.; Tschampel, S. M.; Woods, R. J.; Engelsen, S. *J. Comput. Chem.* **2004**, *25*, 573.

(80) Gonzalez-Outeirino, J.; Kadirvelraj, R.; Woods, R. J. *Carbohydr. Res.* **2005**, *340*, 1007.

(81) Gonzalez-Outeirino, J.; Kirschner, K. N.; Thobhani, S.; Woods, R. J. *Can. J. Chem.* **2006**, *84*, 569.

(82) Woods, R. J. *Glycoconjugate J.* **1998**, *15*, 209.

(83) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208.

(84) Gao, J. L.; Habibollazadeh, D.; Shao, L. *J. Phys. Chem.* **1995**, *99*, 16460.

(85) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.

(86) Noskov, S. Y.; Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2005**, *109*, 6705.

(87) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.

(88) Yu, H.; Geerke, D. P.; Liu, H.; van Gunsteren, W. E. *J. Comput. Chem.* **2006**, *27*, 1494.

(89) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2007**, *3*, 1927.

(90) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D. *J. Chem. Theory Comput.* **2007**, *3*, 1120.

(91) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549.

CT100450S

JCTC Journal of Chemical Theory and Computation

A Transferable Nonbonded Pairwise Force Field to Model Zinc Interactions in Metalloproteins

Ruibo Wu,[†,‡] Zhenyu Lu,[†] Zexing Cao,[‡] and Yingkai Zhang*[,†]

Department of Chemistry, New York University, New York, New York 10003, United States and Department of Chemistry and State Key Laboratory of Physical Chemistry of Solid Surfaces, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China

Received September 15, 2010

Abstract: Herein we introduce a novel practical strategy to overcome the well-known challenge of modeling the divalent zinc cation in metalloproteins. The main idea is to design short−long effective functions (SLEF) to describe charge interactions between the zinc ion and all other atoms. This SLEF approach has the following desired features: (1) It is pairwise, additive, and compatible with widely used atomic pairwise force fields for modeling biomolecules; (2) It only changes interactions between the zinc ion and other atoms and does not affect force field parameters that model other interactions in the system; (3) It is a nonbonded model that is inherently capable to describe different zinc ligands and coordination modes. By optimizing two SLEF parameters as well as zinc van der Waals parameters through force matching based on Born−Oppenheimer ab initio quantum mechanical/molecular mechanical (QM/MM) molecular dynamics (MD) simulations, we have successfully developed the first SLEF force field (SLEF1) to describe zinc interactions. Extensive MD simulations of seven zinc enzyme systems with different coordination ligands and distinct chelation modes (four-, five-, and six-fold), including a binuclear zinc active site, yielded zinc coordination numbers and binding distances in good agreement with the corresponding crystal structures as well as ab initio QM/MM MD results. This not only demonstrates the transferability and adequacy of the new SLEF1 force field in describing a variety of zinc proteins but also indicates that this novel SLEF approach is a promising direction to explore for improving force field description of metal ion interactions.

1. Introduction

Zinc proteins constitute approximately 10% of the total human proteome[1] and play a variety of essential biological roles,[2−5] such as transcription factors, signaling proteins, and transport/storage proteins as well as enzymes. Their function and/or structural organization are critically dependent on the zinc binding site,[4,6−8] which can be classified as catalytic, structural, inhibitory, and protein interface zinc sites based on the role of the divalent zinc cation. Typical zinc ligands include side chains of Cys, His, Glu, and Asp, water molecules, and other small molecules. A key feature of the zinc coordination is its flexibility:[4,9−12] it can adopt multiple binding modes, including tetrahedral-, penta-, or hexacoordination geometry. Especially for the zinc coordination to the carboxylate group, it could be either bidentate or monodentate. This inherent flexibility of zinc coordination poses a daunting challenge for all currently available pairwise atomic force fields to describe zinc interactions,[13−18] including bonded,[19−24] nonbonded,[25] and semibonded[26] models.

In the bonded model,[19−24] zinc−ligand coordination interactions are modeled as covalent bonds, and the desired zinc coordination geometry is maintained by employing explicit bonding and angle bending terms. This clearly prevents any change of the zinc coordination mode or ligand exchange, therefore not suitable for describing the dynamics of zinc coordination. For the nonbonded model,[25] in which

* Corresponding author. E-mail: yingkai.zhang@nyu.edu.
[†] New York University.
[‡] Xiamen University.

10.1021/ct100525r
Published on Web 12/23/2010

interactions between the zinc ion and all other atoms are described by electrostatics and van der Waals (vdW) terms, it has been notoriously known for its failure in describing the tetra- or pentacoordinated zinc cation.[13−18] Previous simulations of several zinc-containing proteins with nonbonded models have led to very different coordination modes in comparison with corresponding X-ray structures.[13−15,17] In the semibonded model,[26,27] virtual fractional charges around a metal atom are employed to mimic valence electrons. It has been shown to describe the tetracoordinated zinc ion well, but its capability to model penta- and hexacoordination has not been demonstrated. Currently, it has been widely thought that pairwise atomic force fields may be inherently unsuitable for describing flexible zinc coordination, and it would be necessary to employ polarizable force fields or quantum mechanical/molecular mechanical (QM/MM) methods to explicitly take account of polarization and charge-transfer effects between Zn^{2+} and its ligands.[13,17,28−33]

In this work, we are motivated to develop a novel practical strategy to tackle this well-known challenge of modeling the divalent zinc cation in metalloproteins. The working hypothesis is that the main deficiency of existing nonbonded models comes from the $1/r$ function form for the charge−charge interaction term. It is not appropriate to describe zinc coordination bonding, although it may be reasonable to describe long-range electrostatic interactions between the Zn^{2+} ion and other atoms beyond the first coordination shell. Thus our main idea is to design short−long effective functions (SLEF) to describe charge interactions between the zinc ion and all other atoms. The short-range is designed to describe the coordination bonding between the zinc ion and its ligands, while the other behaves similar to $1/r$ for long-range electrostatic interactions. Herein by optimizing a total of four parameters through force matching[34−37] based on Born−Oppenheimer ab initio QM/MM molecular dynamics (MD) simulations,[10,38−43] we have successfully developed the first SLEF force field (SLEF1) to describe zinc interactions compatible with the amber99SB force field[44−46] and the TIP3P[47] water model and demonstrated its good transferability and adequacy in describing a variety of zinc proteins.

2. Methods

A. Nonbonded SLEF Force Field to Model Zinc Interactions. In the current work, we have introduced the following novel short−long effective function (SLEF) to describe charge interactions between a divalent zinc ion i and any other atom j:

$$E_{\mathrm{es,SLEF}}^{\mathrm{Zn},j}(r_{i,j}) = \frac{1}{4\pi\varepsilon_0}\left\{\frac{q_{\mathrm{Zn}}q_j}{\sqrt{r_{i,j}^2 + \alpha \times \frac{q_j^2}{(R_i^* + R_j^*)} \times \exp(\beta \times r_{i,j}^2)}} + \frac{1}{1 + \exp\left(-2\left(\frac{2r_{i,j}}{3} - 1.0\right)\right)} \times \frac{q_{\mathrm{Zn}}q_j}{r_{i,j}}\right\} \quad (1)$$

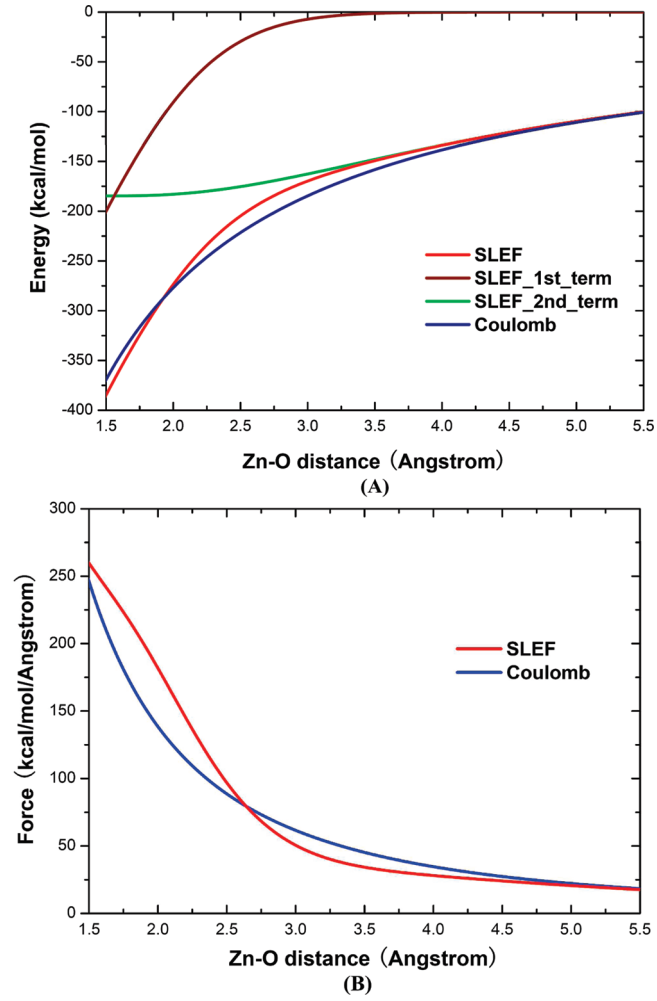


Figure 1. Illustration of the difference between SLEF and the conventional $1/r$ Coulomb function in describing charge interactions between Zn^{2+} and the oxygen of TIP3P water: (A) energy and (B) force. The parameters in the SLEF1 force field were employed.

where $r_{i,j}$ is the distance, q_{Zn} refers the charge of the zinc ion which has a value of 2.0, q_j is the MM charge of the atom j, R^* refers to the vdW radii, and α and β are two new positive parameters which need to be determined. As shown in Figure 1, the first term only makes a contribution at the short range, while the second term employing a similar damping function used in DFT dispersion correction approach[48,49] is relatively flat in the short range but turns into $1/r$ in the long range (>4.5 Å). Thus the main difference between our introduced SLEF function and the coulomb function form $1/r$ is at the short-range, where the coordination interaction is expected to be dominant.

Besides the charge-interaction term, the conventional Lennard-Jones 12−6 function form has been employed to describe the vdW interactions between a zinc ion i and any other atom j:

$$E_{\mathrm{vdW}}^{\mathrm{Zn},j}(r_{i,j}) = \varepsilon_{ij}\{(R^*_{i,j}/r_{i,j})^{12} - 2(R^*_{i,j}/r_{i,j})^6\}, \quad R^*_{i,j} = R^*_i + R^*_j, \quad \varepsilon_{ij} = \sqrt{\varepsilon_i\varepsilon_j} \quad (2)$$

We can see that the above SLEF approach to describe zinc interactions has the following desired features: (1) It is

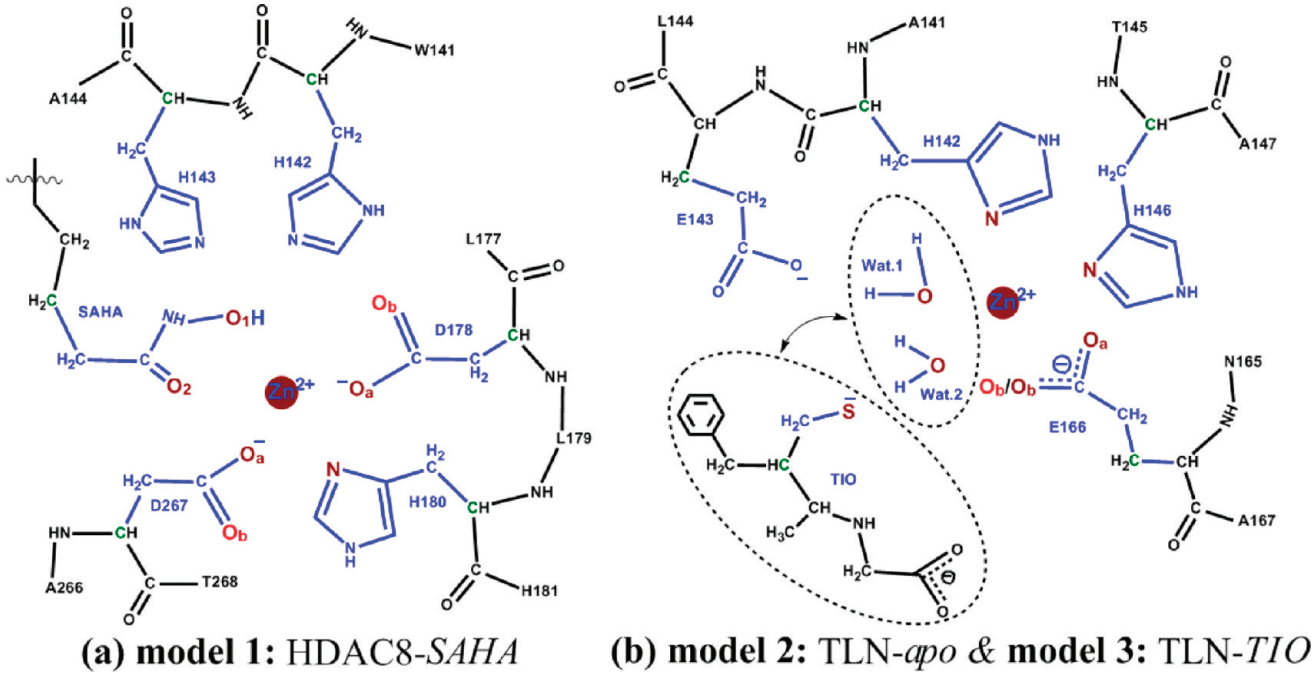


Figure 2. Illustration of various coordination shells in three zinc enzyme systems in our training set used for optimizing SLEF1 force field parameters. Those atoms selected for force matching include the zinc cation and directly/potentially coordinated atoms (colored in brown/red).

pairwise, additive, and compatible with widely used atomic pairwise force fields for modeling biomolecules; (2) It only changes interactions between the zinc ion and other atoms and does not affect force field parameters that model other interactions in the system; and (3) It is a nonbonded model that is inherently capable to describe different zinc ligands and coordination modes. Thus to extend the widely employed atomic pairwise force fields to simulate zinc metalloproteins with the SLEF approach, it only needs to determine four additional parameters: α and β in the SLEF function (eq 1) and two zinc vdW parameters: ε and R^* (eq 2).

B. Parameterization with Force Matching Based on Ab initio QM/MM MD Simulations. Force matching (FM)[34−37] has become a powerful and increasingly popular approach to parametrize atomic force fields based on high-level quantum mechanical calculations. Here we have adapted the ab initio QM/MM force matching approach[36,37] to determine the four parameters (two vdW parameters for Zn, ε and R^* and two parameters in SLEF function, α and β) by minimizing the following target function:

$$\chi^2 = \sum_I \sum_J \sum_k ||f_{I,J,k}^{\mathrm{SLEF}} - F_{I,J,k}^{\mathrm{ref}}||^2$$

where $F_{I,J,k}^{\mathrm{ref}}$ refers the reference force from ab initio QM/MM calculations on the k atom with the J_{th} configuration of the I enzyme system in the training set, and $f_{I,J,k}^{\mathrm{SLEF}}$ is the corresponding force calculated based on the SLEF force field.

In the current study, our training set consists of three zinc enzyme systems: HDAC8-SAHA,[50] TLN-apo,[51] and TLN-TIO,[52] as shown in Figure 2. These three systems represent five-, six-, and four-fold zinc coordination, respectively, and the ligands are typical in zinc proteins: His, Glu/Asp, Cys, hydroxamate, and water. All chosen configurations are snapshots from Born−Oppenheimer ab initio QM/MM MD simulations, as described in detail in our previous work.[10] For each enzyme system, 25 ps B3LYP(SDD,[53]6-31G*) QM/MM MD simulations had been carried out, and 200 snapshots from the last 20 ps have been chosen for force matching. This level of QM treatment has been extensively tested and employed successfully to describe the zinc coordination shell[10,43,54−57] and is similar to other recent ab initio QM/MM studies of zinc enzymes.[58,59] A total of 600 configurations have been employed in parametrization with the amber99SB force field[44−46] for modeling proteins and the TIP3P[47] water model. For each configuration, the reference forces on selected atoms, including the zinc cation, all directly and potentially coordinated atoms (illustrated in Figure 2), have been calculated by performing two B3LYP-(SDD,6-31G*) QM/MM calculations. One calculation is on the whole system, and the other is on the same system without the zinc ion. The force difference between two calculations can be considered as the force coming from its interaction with zinc and has been employed as the reference force $F_{I,J,k}^{\mathrm{ref}}$. Correspondingly, the $f_{I,J,k}^{\mathrm{SLEF}}$ is calculated with the SLEF1 force field. The advantage of employing this force difference is that the parametrization of the four parameters would be solely dependent on zinc interactions, which is and should be much desired. In addition, such an ab initio QM/MM MD force matching approach allows us to employ a large amount of information from a first principle description of zinc interactions, while properly taking account of the heterogeneous enzyme environment and the dynamic fluctuations. All ab initio QM/MM calculations were performed with modified Q-Chem[60] and Tinker[61] programs, and the QM/MM boundaries were described by the pseudobond approach[62−65] with the improved parameters.[62]

Table 1. Resulting Four Parameters of the SLEF1 Force Field to Model Zinc Interactions[a]

α	β	R^*	ε
2.23	1.04	1.21	0.23

[a] Units: α, Å^3/e^2; β, 1.04 Å^{-2}; R^*, Å; and ε, kcal/mol; α and β are parameters in the SLEF function (eq 1) and R^* and ε are vdW parameters of zinc.

The four parameters, including two vdW parameters ε and R^* for zinc and two parameters α and β in the SLEF function (eq 1), were determined by the parameter scan combined with local minimization procedure to effectively explore the parameter space. Specifically, ε value has been scanned from 0.05 to 0.50 with 0.01 step size and the other three parameters are optimized at each scan step. The simplex algorithm[66] implemented in GNU Scientific Library (GSL) and the modified Tinker program[61] were employed in the parametrization procedure. The resulting four parameters for this new SLEF1 force field describing the zinc interactions compatible with the amber99SB force field[44−46] and the TIP3P[47] water model are listed in Table 1.

C. Tests. We have implemented the new SLEF1 force field in the modified Tinker program.[61] In order to examine its transferability and performance, we have carried out extensive MD simulations of seven zinc enzyme complexes with different coordination ligands (Asp/Glu, His, Cys, water, and small molecules) and distinct chelation modes (four-, five-, and six-fold), including the binuclear zinc active site. Besides three systems in the training set, as illustrated in Figure 2, the four additional models are: (A) an HDAC8-substrate complex system[67] which has a five-fold coordinated zinc catalytic site; (B) an HDAC7-SAHA complex[68] which has two four-fold coordinated zinc binding sites, one catalytic site and one Cys-rich structural site; (C) a carbonic anhydrase (CAII) enzyme system[69] which has a tetrahedral coordinated zinc catalytic site; and (D) an L-rhamnose isomerase enzyme[70] containing a binuclear zinc coordination shell.

For each enzyme system, the initial structure was prepared based on the corresponding crystal structure.[67−70] Then 4 ns MD simulations with the SLEF1 force field describing zinc interactions were carried out at 300 K with a time step of 1 fs. Amber99SB force field[44−46] was used for protein residues, TIP3P model[47] for water molecules, and generalized AMBER force field (GAFF)[71] for the other small molecules. The 18 and 12 Å cutoffs were employed for electrostatic and vdW interactions. For comparison, 4 ns MD simulations with the conventional nonbonded zinc model[25] (called as the Coulomb scheme) have also been performed.

3. Results

A. Performance of the SLEF1 Force Field on Three Zinc Enzymes in the Training Set. By optimizing two SLEF parameters as well as zinc vdW parameters through force matching based on ab initio QM/MM MD simulations, we have successfully developed the first SLEF force field to describe zinc interactions compatible with the amber99SB force field and the TIP3P water model. The four parameters are presented in Table 1, with α and β as 2.23 Å^3/e^2 and 1.04 Å^{-2}, respectively, and the vdW parameters of Zn are R^*=1.21 Å; ε=0.23 kcal/mol.

The force errors on the selected ligand atoms from the SLEF1 force field as well as other MM models and QM/MM calculations with different basis sets (denoted as "DBS") for three zinc enzymes in the training set are summarized in Tables 2−4. Not surprisingly, the Coulomb scheme, in which Stote's parameters[25] for zinc and Amber99SB force field for other atoms were used, gives the largest force errors for each model. For the vdW FM scheme, in which vdW parameters of selected atoms (Zn and four kinds of ligand−atom: His−N, water−O, Glu/Asp−O, S) were optimized by FM,

Table 2. Force Error Calculated for the HDAC8-SAHA System with a Pentacoordinated Zinc Binding Site[a]

	rms force error (kcal/mol/Å)				
		ligand atom (the force derived from Zn)			
HDAC8-SAHA(model 1)	Zn (total force)	H180 (N)	D178 (O_a/O_b)	D267 (O_a/O_b)	SAHA (O_1/O_2)
Coulomb + LJ-$R_{12,6}$ (Zn, Stote) [b] (Coulomb scheme)	57.8	11.8	56.7/18.3	30.5/18.2	34.7/21.2
Coulomb + LJ- $R_{12,6}$ (Zn+Ligands, FM) [c] (vdW FM scheme)	48.7	14.5	19.2/15.7	50.1/21.3	19.1/17.1
SLEF(α; β) + LJ- $R_{12,6}$ (Zn, FM) [d] (SLEF scheme)	**23.2**	**9.5**	**24.7/10.4**	**11.1/7.8**	**14.1/9.1**
Different Basie Set in QM/MM [e] (DBS)	3.4	4.0	2.4/2.5	2.6/2.0	5.0/1.8

[a] The reference forces are calculated with B3LYP(SDD,6-31G*) QM/MM calculations. [b] Stote's vdW parameters[25] for the zinc ion: R^* = 1.09 and ε = 0.25. [c] Used the Coulomb function to describing charge interactions; vdW parameters of the zinc ion and the 4 types of coordinated atoms (a total of 10) were optimized by force matching. [d] Using the developed SLEF1 force field. (a total of 4 parameters have been optimized: α = 2.23; β = 1.04; σ =1.21; and ε = 0.23;). [e] DBS indicates the force difference derived from using different basis sets (DBS) in QM/MM calculations (level 1: SDD for zinc, other atoms by 6-31G*;and level 2: 6-311G** for all atoms).

Table 3. Force Error Calculated for the TLN-Apo System with a Hexacoordinated Zinc Binding Site[a]

	rms force error (kcal/mol/Å)					
		ligand atom (the force derived from Zn)				
TLN-apo (model 2)	Zn (total force)	H142 (N)	H146 (N)	E166 (O_a/O_a)	water1 (O)	water2 (O)
Coulomb + LJ-$R_{12,6}$ (Zn, Stote)[b] (Coulomb scheme)	29.3	10.1	22.3	30.9/39.4	57.1	56.2
Coulomb + LJ- $R_{12,6}$ (Zn+Ligands, FM)[c] (vdW FM scheme)	15.9	5.7	10.6	19.0/5.1	16.7	15.7
SLEF(α; β) + LJ- $R_{12,6}$ (Zn, FM)[d] (SLEF scheme)	**18.4**	**6.7**	**12.4**	**9.3/12.6**	**18.7**	**18.3**
Different Basie Set in QM/MM[e] (DBS)	2.5	3.9	3.6	1.7/0.9	4.0	3.8

[a] The reference forces are calculated with B3LYP(SDD,6-31G*) QM/MM calculations. For other descriptions see Table 2.

Table 4. Force Error Calculated for the TLN-TIO System with a Tetracoordinated Zinc Binding Site[a]

	rms force error (kcal/mol/Å)				
		ligand atom (the force derived from Zn)			
TLN-TIO (model 3)	Zn (total force)	H142 (N)	H146 (N)	E166 (O_a/O_b)	TIO (S)
Coulomb + LJ-$R_{12,6}$ (Zn, Stote)[b] (Coulomb scheme)	47.3	18.6	13.6	35.5/15.2	43.6
Coulomb + LJ- $R_{12,6}$ (Zn+Ligands, FM)[c] (vdW FM scheme)	29.6	13.2	10.6	37.5/14.7	26.8
SLEF(α; β) + LJ- $R_{12,6}$ (Zn, FM)[d] (SLEF scheme)	**26.1**	**10.3**	**7.9**	**13.3/11.0**	**9.5**
Different Basie Set in QM/MM[e] (DBS)	11.5	5.6	5.5	8.1/3.2	10.6

[a] The reference forces are calculated with B3LYP(SDD,6-31G*) QM/MM calculations. For other descriptions see Table 2.

the force errors are reduced for all models. The SLEF scheme gives the smallest force errors overall among the three MM schemes. It should be noted that there are a total of 10 parameters optimized in the vdW FM scheme, while only 4 parameters optimized in the SLEF1 force field. As shown in Tables 2 and 4, the force errors of several reference atoms in the vdW FM scheme are significantly larger than those in the SLEF1 force field, such as those of Zn and D267(O_a/O_b) in the model 1 which has a five-fold zinc coordination and E166 (O_a/O_b) and TIO (S) in the model 3 which has a four-fold zinc coordination. Therefore, the SLEF function plays an important contribution to decrease the force errors for four- and five-fold zinc coordination shells. As a result, we found that MD simulations with the vdW FM scheme could not reproduce the similar zinc coordination as observed in crystal structures and in ab initio QM/MM MD simulations for models 1 and 3, while the SLEF1 force field yields good results in MD simulations of all three systems, as shown in Figure 3. These results lend further support for our working hypothesis that the difficulty of the conventional nonbonded zinc model in describing zinc-coordination may come from the $1/r$ function form for the charge interaction term. Meanwhile, we can see that the error with the SLEF1 force field is still significantly larger than the DBS error, which indicates that there is significant room to further improving the description of zinc interactions.

The test results of the amber99SB-SLEF1 force field in describing the zinc coordination shell for three enzymes in the training set are presented in Figure 3. We can see that for all three systems, simulations with the amber99SB-SLEF1 force field yield zinc coordination geometries consistent with both experimentally determined X-ray structures and ab initio QM/MM MD simulation results. On the other hand, for the conventional nonbonded model with the $1/r$ form for zinc charge interactions, it yields very different coordination geometries. Meanwhile, with the corresponding crystal structure as the reference, we can see that the root-mean-square deviation (rmsd) of heavy atoms in the first zinc coordination shell is significantly smaller for simulations with the amber99SB-SLEF1 force field.

Model 1 (HDAC8-SAHA System). As shown in Figure 3, a five-fold coordination geometry is observed in the crystal structure[50] and has been reproduced well by ab initio QM/MM MD simulations.[10] A similar coordination structure as well as important hydrogen-bond interactions between SAHA and His142/His143 are also well maintained in our simulations with the new SLEF1 force filed. But for the conventional Coulomb scheme, it yields a six-fold coordination geometry for zinc and significantly changes the active site geometry. Due to the wrong bidentate chelation of the two Asp residues, O_2 of SAHA is no longer bonded with zinc in simulations with the conventional nonbonded zinc model, and there is no hydrogen bond between SAHA and His142.

Model 2 (TLN-apo System). The six-fold zinc coordination structure[51] was reproduced well in our previous QM/MM simulations and with the SLEF scheme. But for simulations with the conventional Coulomb scheme, the coordination interaction between His142 and Zn was replaced by another water molecule instead, and its Zn coordination shell is much different from the crystal structure, which is also demonstrated by the rmsd curve. It should be noted that the flexible behavior of Glu166 observed in our ab initio QM/MM MD simulations was not be observed in our simulations with the amber99-SLEF1 force field, which indicates its limitation.

Model 3 (TLN-TIO System). Both ab initio QM/MM MD simulations[10] and the SLEF scheme yielded a similar tetrahedral coordination geometry as in the crystal structure.[52] But the Glu166 was bidentate with zinc in the Coulomb scheme, resulting in a five-fold zinc coordination geometry. Meanwhile, the Zn−S coordination bond, which is obviously too short in the Coulomb scheme (2.02 Å), is significantly improved in simulations with the SLEF1 force field (2.24 Å).

B. Tests of the Transferability of SLEF1 on Other Zinc Enzyme Models. To further test the transferability and the performance of the resulting amber99SB-SLEF1 force field, we have further carried out MD simulations on four additional zinc enzyme systems including: Model A, a HDAC8−substrate complex system which has a five-fold coordinated zinc catalytic site; Model B, a HDAC7−SAHA complex which has two four-fold coordinated zinc binding sites—one catalytic and one Cys-rich structural site; Model C, a carbonic anhydrase (CAII) enzyme system which has a four-fold coordinated zinc catalytic site; and Model D, a L-rhamnose isomerase enzyme containing a binuclear zinc coordination site. The results are presented in Figures 4−7. We can see that simulations with the SLEF scheme yield zinc coordination geometries consistent with crystal structures and are significantly better than those results from MD simulations using the conventional Coulomb scheme. The various coordination numbers observed in crystal structures are well maintained in simulations with the SLEF scheme, while the Coulomb scheme tends to yield higher coordination numbers for most zinc coordination shells.

Model A (HDAC8−Substrate). As shown in Figure 4, Model A is a substrate-bound HDAC8 system. Our previous QM/MM simulations[43] yielded a five-fold zinc coordination shell, consistent with the observations from the crystal

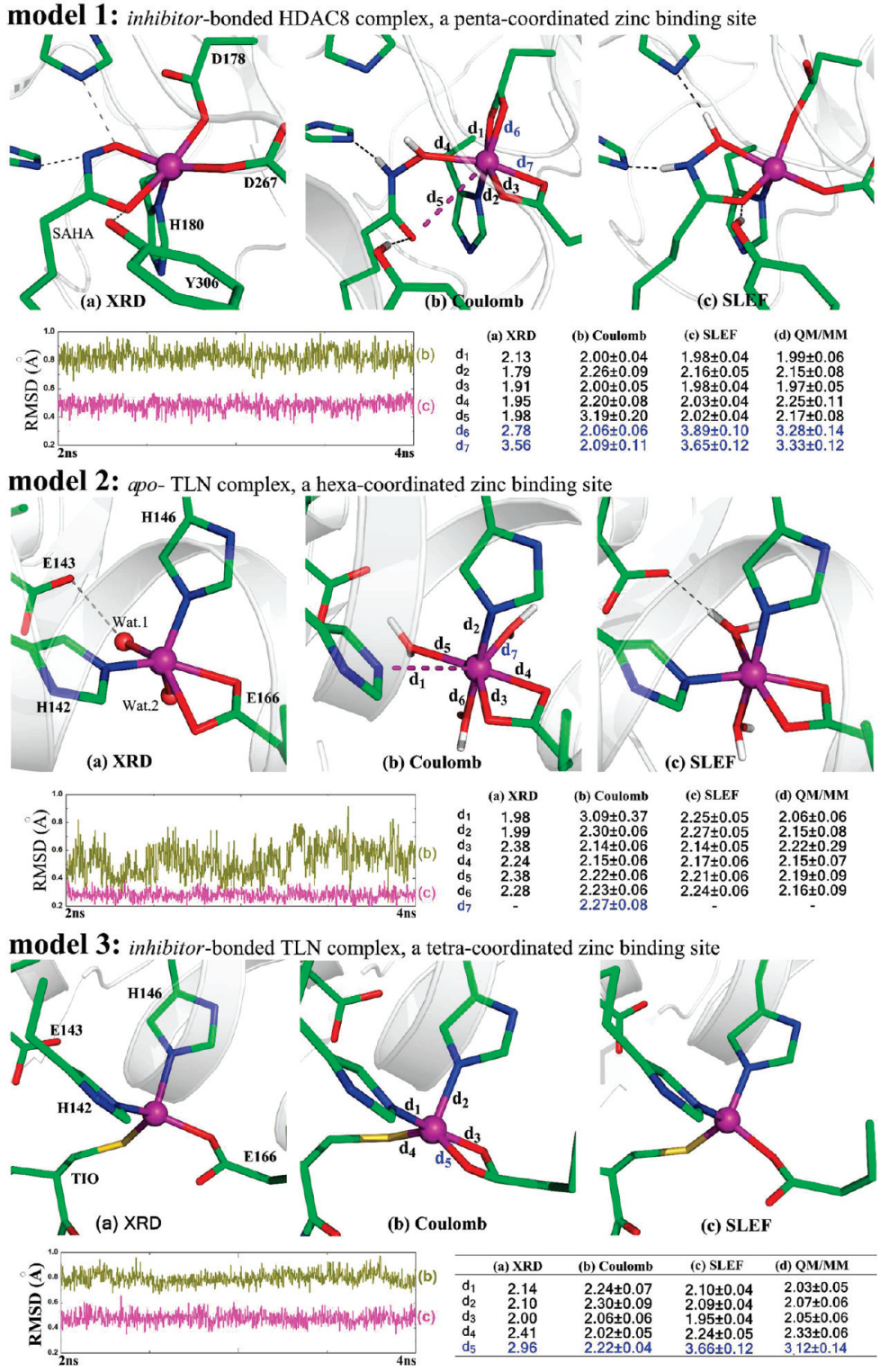
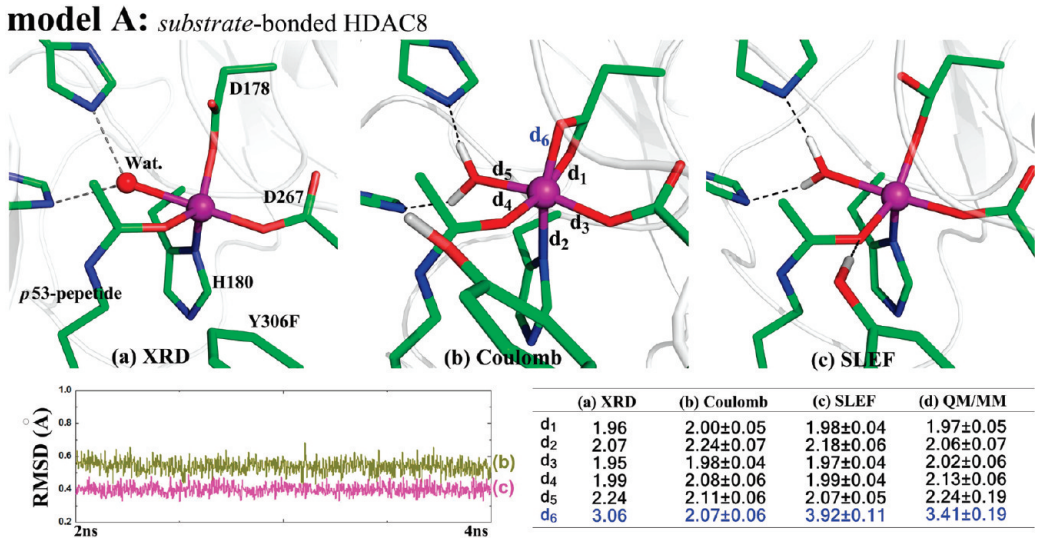


Figure 3. Test results on the three zinc enzyme systems in the training set. XRD refers to results in crystal structures;[50−52] Coulomb refers to results calculated from 4 ns MD simulations with the amber99SB force field and the nonbonded Coulomb model for zinc; and SLEF denotes results from 4 ns MD simulations with the amber99SB force field and our parametrized SLEF1 model for zinc interactions. QM/MM indicates the results from 25 ps B3LYP(SDD, 6-31G*) QM/MM MD simulations.[10]

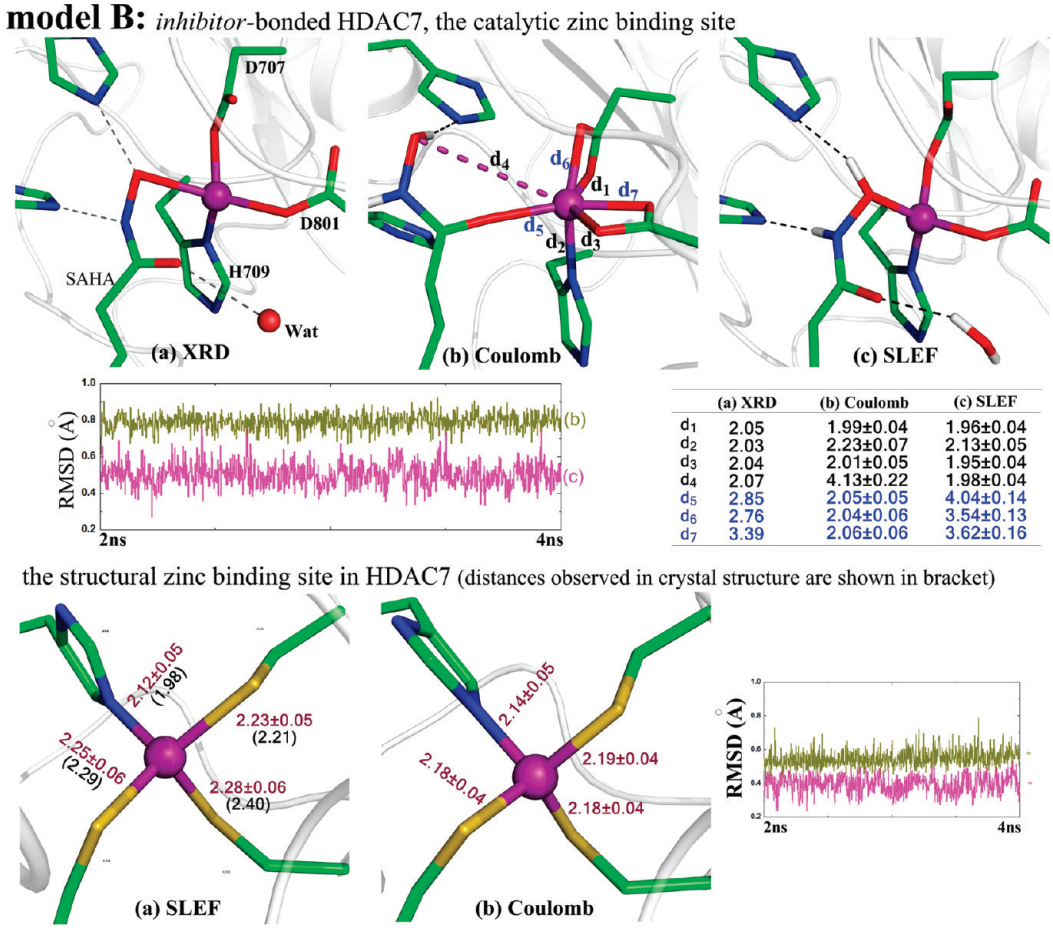
structure.[67] With the SLEF scheme, both the coordination number and the important hydrogen-bonds around the p-53 peptide substrate are kept very well during the MD simulation. On the other hand, for simulations with the Coulomb scheme, although the rmsd value is also small, the penta-coordinated structure is not maintained due to the bidentate chelation of Asp178, and there is no hydrogen bond between Y306 and the p-53 peptide.

Model B (HDAC7−SAHA). In comparison with the Model 1 (HDAC8−SAHA) complex in the training set, the coordination residues and inhibitors are the same, but different SAHA−zinc chelation modes have been observed in crystal



	(a) XRD	(b) Coulomb	(c) SLEF	(d) QM/MM
d_1	1.96	2.00±0.05	1.98±0.04	1.97±0.05
d_2	2.07	2.24±0.07	2.18±0.06	2.06±0.07
d_3	1.95	1.98±0.04	1.97±0.04	2.02±0.06
d_4	1.99	2.08±0.06	1.99±0.04	2.13±0.06
d_5	2.24	2.11±0.06	2.07±0.05	2.24±0.19
d_6	3.06	2.07±0.06	3.92±0.11	3.41±0.19

Figure 4. Test results on the Model A system. XRD refers to the crystal structure.[67] QM/MM indicates the results from 25 ps B3LYP(SDD, 6-31G*) QM/MM MD simulations.[43] For other descriptions see Figure 3.

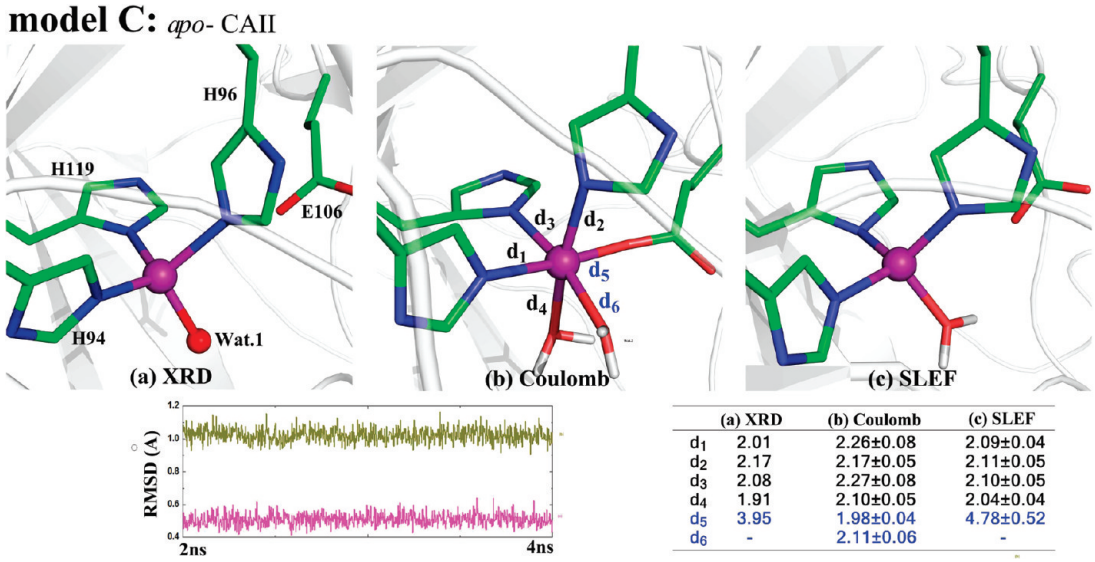


	(a) XRD	(b) Coulomb	(c) SLEF
d_1	2.05	1.99±0.04	1.96±0.04
d_2	2.03	2.23±0.07	2.13±0.05
d_3	2.04	2.01±0.05	1.95±0.04
d_4	2.07	4.13±0.22	1.98±0.04
d_5	2.85	2.05±0.05	4.04±0.14
d_6	2.76	2.04±0.06	3.54±0.13
d_7	3.39	2.06±0.06	3.62±0.16

Figure 5. Test results on the Model B system. XRD refers to the crystal structure.[68] For other descriptions see Figure 3.

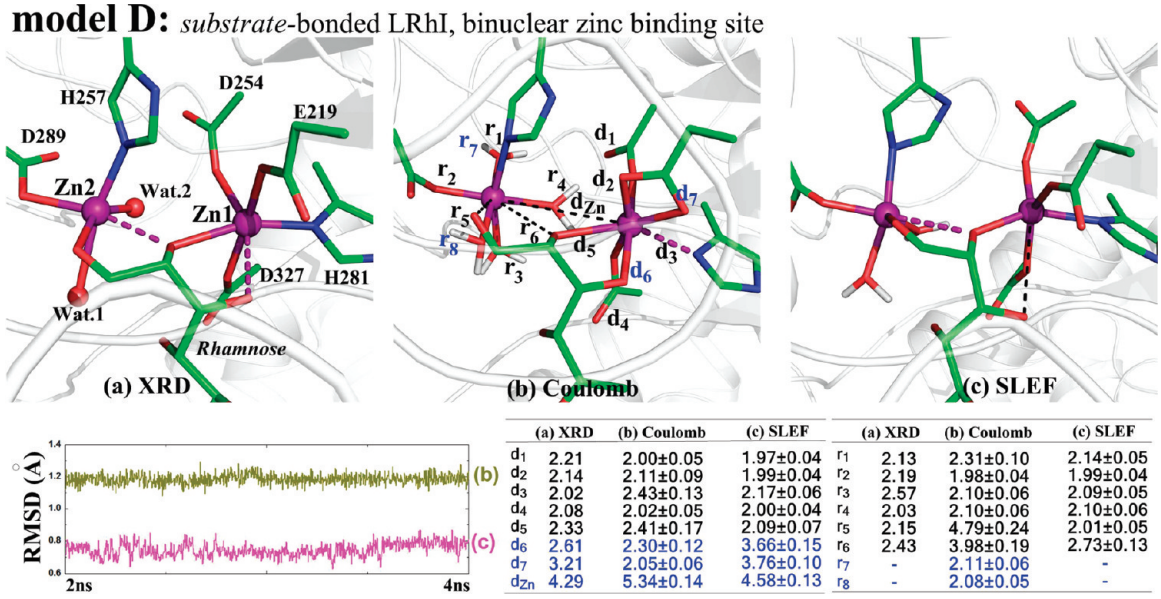
structures: monodentate and a four-fold zinc coordination shell in HDAC7,[68] while bidentate and a five-fold coordination in HDAC8.[50] Such a distinct coordination mode has also been confirmed by ab initio QM/MM MD simulations and thus serves as a stringent test for the force field description. From Figure 5, we can see that MD simulations with the SLEF scheme yield a zinc coordination shell consistent with the X-ray structure and ab initio QM/MM MD simulations, maintain the important hydrogen-bond network, and show a smaller rmsd value. On the other hand, the conventional Coulomb scheme leads to a six-fold coordinated structure.

Besides the catalytic zinc site, there is another Cys-rich structural zinc coordination motif in HDAC7. The SLEF scheme can describe this coordination shell very well, as shown in Figure 5. Although the Coulomb scheme can also



	(a) XRD	(b) Coulomb	(c) SLEF
d_1	2.01	2.26±0.08	2.09±0.04
d_2	2.17	2.17±0.05	2.11±0.05
d_3	2.08	2.27±0.08	2.10±0.05
d_4	1.91	2.10±0.05	2.04±0.04
d_5	3.95	1.98±0.04	4.78±0.52
d_6	-	2.11±0.06	-

Figure 6. Test results on the Model C system. XRD refers to the crystal structure.[69] For other descriptions see Figure 3.



	(a) XRD	(b) Coulomb	(c) SLEF
d_1	2.21	2.00±0.05	1.97±0.04
d_2	2.14	2.11±0.09	1.99±0.04
d_3	2.02	2.43±0.13	2.17±0.06
d_4	2.08	2.02±0.05	2.00±0.04
d_5	2.33	2.41±0.17	2.09±0.07
d_6	2.61	2.30±0.12	3.66±0.15
d_7	3.21	2.05±0.06	3.76±0.10
d_{Zn}	4.29	5.34±0.14	4.58±0.13

	(a) XRD	(b) Coulomb	(c) SLEF
r_1	2.13	2.31±0.10	2.14±0.05
r_2	2.19	1.98±0.04	1.99±0.04
r_3	2.57	2.10±0.06	2.09±0.05
r_4	2.03	2.10±0.06	2.10±0.06
r_5	2.15	4.79±0.24	2.01±0.05
r_6	2.43	3.98±0.19	2.73±0.13
r_7	-	2.11±0.06	-
r_8	-	2.08±0.05	-

Figure 7. Test results on the Model D system. XRD refers to the crystal structure.[70] For other descriptions see Figure 3.

obtain the correct coordination number, the Zn−S coordination distance is very short.

Model C (CAII−Apo). The apo structure of carbonic anhydrase (CAII) has a tetrahedral zinc active site,[69] as shown in Figure 6. Herein our SLEF scheme also reproduces the four-fold coordination shell very well, but the Coulomb scheme leads to the hexacoordination. It seems that Coulomb scheme overestimates the electrostatic interaction between E106 and the divalent zinc cation, which is ∼4 Å apart in the crystal structure.[69]

Model D (L-RhI). L-Rhamnose isomerase, which can efficiently catalyze the isomerization between various aldoses and ketoses, has a binuclear zinc coordination shell.[70] Although no such binuclear zinc active site has been employed in the parameter optimization, the resulted SLEF1 force field can describe this challenging case[72] relatively well, including the zinc−zinc distance. As seen in Figure 7, it improved significantly against the conventional coulomb scheme in terms of both the zinc coordination spheres and the rmsd from the crystal structure. Meanwhile, these test results indicate that the SLEF1 force field still needs to be further improved, and a binuclear zinc active site should also be included in the training set in the future development.

4. Discussion

The above tests clearly demonstrated that the conventional coulomb scheme has two main deficiencies in describing zinc coordinations: (1) Its strong preference of water coordination and bidentate chelation of Asp/Glu residues leads to higher coordination numbers for most zinc coordination shells; and (2) its coordination to the neutral His residue can be substituted by a water molecule or a carboxylate ligand, such as in Model 2. Both deficiencies have been overcome by our developed SLEF1 force field, which has yielded zinc coordination structures in very good agreement with the corresponding crystal structures as well as ab initio QM/

Table 5. Mean and Maximum Deviation between the Coordination Distances in Crystal Structures and those from SLEF1 MD Simulations[a]

	mean deviation (Å)	maximum deviation(Å)
negative ligands	−0.09	+0.07/−0.24
neutral ligands	0.04	+0.41/−0.48

[a] The coordination distances from the SLEF1 MD simulations are the averaged values from the MD trajectories. The coordination distances in seven zinc enzyme complexes (training and test models) are all considered.

MM MD results. As summarized in Table 5, the mean deviation between the coordination distances in crystal structures and the corresponding average MD value from our SLEF1 simulations is 0.04 and −0.09 Å, respectively, and the largest deviation is +0.41/−0.48 Å, which is from the binuclear zinc binding site of L-RhI. For all seven zinc enzyme complexes, the coordination modes observed in crystal structures are well reproduced in simulations with the SLEF1 force field. In particular, for HDAC8−SAHA (Model 1) and HDAC7−SAHA (Model B) systems, their coordination ligands are the same, but different coordination modes have been observed in crystal structures: a four-fold zinc coordination shell in HDAC7,[68] while a five-fold coordination in HDAC8.[50] Such two distinct coordination modes with the same coordination ligands have been well reproduced in our simulations with the SLEF1 force field. Meanwhile, this would pose a fundamental challenge for bonded models[19−24] to describe zinc interactions, which usually assumes that the same set of coordination ligands would adopt the same coordination mode.

In comparison to the conventional Coulomb function $1/r$, the key difference of our introduced SLEF function is in the short range (<4.5 Å), where the coordination interaction is expected to be dominant. From Figure 1, we can see that the resulted energy and force from the SLEF approach do not parallel those from the Coulomb function in the short-range regime. Meanwhile, the difference of the SLEF function to $1/r$ is varied with the magnitude of charges since the charge also appears in the denominator of the short-range function in eq 1. Thus, the SLEF function cannot be considered as simply scaling the charge in the short-range regime and then returning to the $1/r$ form in the long range.

Since the SLEF approach is a nonbonded model to model zinc interactions and only changes interactions between the zinc ion and other atoms, it would be quite straightforward to implement it into typical MD simulation packages: the replacement of the Coulomb function with the SLEF function to describe charge interactions between the zinc ion and all other atoms, and the employment of vdW parameters developed here for the zinc ion.

5. Conclusion

In this work, we have introduced a novel practical strategy to meet the challenge of describing zinc interactions: the design of new short−long effective functions (SLEF) to treat charge interactions between the zinc ion and all other atoms. By optimizing a total of four parameters based on ab initio QM/MM MD simulations and force matching, we have developed the first transferable nonbonded pairwise SLEF force field to describe zinc interactions for modeling zinc metalloproteins compatible with the amber99SB force field and the TIP3P water model. We have carried out MD simulations with the amber99SB-SLEF1 force field on seven different enzymes complexes (a total of nine zinc coordination shells), which include four common kinds of ligands (His, Asp/Glu, Cys, and Water) and various coordination numbers (4, 5, or 6). Most simulations yielded zinc coordination numbers and binding distances in very good agreement with the corresponding crystal structures as well as ab initio QM/MM MD results. These very encouraging results indicate that this novel SLEF approach is a promising and attractive direction to explore for further improving force field description of metalloproteins.

Acknowledgment. This work was supported by NIH (R01-GM079223), NSF (CHE-CAREER-0448156), and the China Scholarship Council. We thank NCSA and NYU-ITS for providing computational resources and Dr. Shenglong Wang for computing support.

References

(1) Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. *J. Proteome. Res.* **2006**, *5*, 196–201.

(2) Parkin, G. *Chem. Rev.* **2004**, *104*, 699–767.

(3) Anzellotti, A. I.; Farrell, N. P. *Chem. Soc. Rev.* **2008**, *37*, 1629–1651.

(4) Maret, W.; Li, Y. *Chem. Rev.* **2009**, *109*, 4682–4707.

(5) Sensi, S. L.; Paoletti, P.; Bush, A. I.; Sekler, I. *Nat. Rev. Neurosci.* **2009**, *10*, 780–791.

(6) Auld, D. S. *Biometals* **2001**, *14*, 271–313.

(7) Lee, Y. M.; Lim, C. *J. Mol. Biol.* **2008**, *379*, 545–553.

(8) Tamames, B.; Sousa, S. F.; Tamames, J.; Fernandes, P. A.; Ramos, M. J. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 466–475.

(9) McCall, K. A.; Huang, C. C.; Fierke, C. A. *J. Nutr.* **2000**, *130*, 1437S–1446S.

(10) Wu, R.; Hu, P.; Wang, S.; Cao, Z.; Zhang, Y. *J. Chem. Theory Comput.* **2010**, *6*, 337–343.

(11) Kuppuraj, G.; Dudev, M.; Lim, C. *J. Phys. Chem. B* **2009**, *113*, 2952–2960.

(12) Babor, M.; Greenblatt, H. M.; Edelman, M.; Sobolev, V. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 221–230.

(13) Banci, L. *Curr. Opin. Chem. Biol.* **2003**, *7*, 143–149.

(14) Donini, O. A. T.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 4180–4188.

(15) Koca, J.; Zhan, C. G.; Rittenhouse, R. C.; Ornstein, R. L. *J. Comput. Chem.* **2003**, *24*, 368–378.

(16) Dal Peraro, M.; Spiegel, K.; Lamoureux, G.; De Vivo, M.; DeGrado, W. F.; Klein, M. L. *J. Struct. Bio.* **2007**, *157*, 444–453.

(17) Zimmer, M. *Coord. Chem. Rev.* **2009**, *253*, 817–826.

(18) Li, X.; Hayik, S. A.; Merz, K. M. *J. Inorg. Biochem.* **2010**, *104*, 512–522.

(19) Vedani, A.; Huhta, D. W. *J. Am. Chem. Soc.* **1990**, *112*, 4759–4767.

(20) Hoops, S. C.; Anderson, K. W.; Merz, K. M. *J. Am. Chem. Soc.* **1991**, *113*, 8262–8270.

(21) Bredenberg, J.; Nilsson, L. *Int. J. Quantum Chem.* **2001**, *83*, 230–244.

(22) Li, W. F.; Zhang, J.; Wang, J.; Wang, W. *J. Am. Chem. Soc.* **2008**, *130*, 892–900.

(23) Lin, F.; Wang, R. X. *J. Chem. Theory Comput.* **2010**, *6*, 1852–1870.

(24) Peters, M. B.; Yang, Y.; Wang, B.; Fusti-Molnar, L.; Weaver, M. N.; Merz, K. M. *J. Chem. Theory Comput.* **2010**, *6*, 2935–2947.

(25) Stote, R. H.; Karplus, M. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, 12–31.

(26) Pang, Y. P. *Proteins: Struct., Funct., Bioinf.* **2001**, *45*, 183–189.

(27) Pang, Y. P.; Xu, K.; El Yazal, J.; Prendergast, F. G. *Protein Sci.* **2000**, *9*, 1857–1865.

(28) Sakharov, D. V.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921–4929.

(29) Sakharov, D. V.; Lim, C. *J. Comput. Chem.* **2009**, *30*, 191–202.

(30) de Courcy, B.; Piquemal, J. P.; Gresh, N. *J. Chem. Theory Comput.* **2008**, *4*, 1659–1668.

(31) Gresh, N.; Piquemal, J. P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113–1130.

(32) Wu, J. C.; Piquemal, J. P.; Chaudret, R.; Reinhardt, P.; Ren, P. Y. *J. Chem. Theory Comput.* **2010**, *6*, 2059–2070.

(33) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565–581.

(34) Ercolessi, F. *Europhys. Lett.* **1994**, *26*, 583–588.

(35) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896–10913.

(36) Akin-Ojo, O.; Song, Y.; Wang, F. *J. Chem. Phys.* **2008**, *129*, 064108.

(37) Maurer, P.; Laio, A.; Hugosson, H. W.; Colombo, M. C.; Rothlisberger, U. *J. Chem. Theory Comput.* **2007**, *3*, 628–639.

(38) Hu, P.; Wang, S.; Zhang, Y. *J. Am. Chem. Soc.* **2008**, *130*, 3806–3813.

(39) Hu, P.; Wang, S.; Zhang, Y. *J. Am. Chem. Soc.* **2008**, *130*, 16721–16728.

(40) Ke, Z.; Zhou, Y.; Hu, P.; Wang, S.; Xie, D.; Zhang, Y. *J. Phys. Chem. B* **2009**, *113*, 12750–12758.

(41) Wang, S.; Hu, P.; Zhang, Y. *J. Phys. Chem. B* **2007**, *111*, 3758–3764.

(42) Zhou, Y.; Wang, S.; Zhang, Y. *J. Phys. Chem. B* **2010**, *114*, 8817–8825.

(43) Wu, R.; Wang, S.; Zhou, N.; Cao, Z.; Zhang, Y. *J. Am. Chem. Soc.* **2010**, *132*, 9471.

(44) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(45) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(46) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

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

(48) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.

(49) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

(50) Somoza, J. R.; Skene, R. J.; Katz, B. A.; Mol, C.; Ho, J. D.; Jennings, A. J.; Luong, C.; Arvai, A.; Buggy, J. J.; Chi, E.; Tang, J.; Sang, B. C.; Verner, E.; Wynands, R.; Leahy, E. M.; Dougan, D. R.; Snell, G.; Navre, M.; Knuth, M. W.; Swanson, R. V.; McRee, D. E.; Tari, L. W. *Structure* **2004**, *12*, 1325–1334.

(51) Holland, D. R.; Hausrath, A. C.; Juers, D.; Matthews, B. W. *Protein Sci.* **1995**, *4*, 1955–1965.

(52) Roderick, S. L.; Fourniezaluski, M. C.; Roques, B. P.; Matthews, B. W. *Biochemistry* **1989**, *28*, 1493–1497.

(53) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866–872.

(54) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Biophys. J.* **2005**, *88*, 483–494.

(55) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Am. Chem. Soc.* **2007**, *129*, 1378–1385.

(56) Xiao, C.; Zhang, Y. *J. Phys. Chem. B* **2007**, *111*, 6229–6235.

(57) Corminboeuf, C.; Hu, P.; Tuckerman, M. E.; Zhang, Y. *J. Am. Chem. Soc.* **2006**, *128*, 4530–4531.

(58) Blumberger, J.; Lamoureux, G.; Klein, M. L. *J. Chem. Theory Comput.* **2007**, *3*, 1837–1850.

(59) Xu, D. G.; Guo, H. *J. Am. Chem. Soc.* **2009**, *131*, 9780–9788.

(60) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J., O. C., Brown, S. T., Gilbert, A. T., Slipchenko, L. V., Levchenko, S. V., O'Neill, D. P., DiStasio, R. A., Lochan, R. C., Wang, T., Beran, G. J., Besley, N. A., Herbert, J. M., Lin, C. Y., Van Voorhis, T., Chien, S. H., Sodt, A., Steele, R. P., Rassolov, V. A., Maslen, P. E., Korambath, P. P., Adamson, R. D., Austin, B., Baker, J., Byrd, E. F., Dachsel, H., Doerksen, R. J., Dreuw, A., Dunietz, B. D., Dutoi, A. D., Furlani, T. R., Gwaltney, S. R., Heyden, A., Hirata, S., Hsu, C. P., Kedziora G., Khalliulin R. Z., Klunzinger, P., Lee, A. M., Lee, M. S., Liang, W., Lotan, I., Nair, N., Peters, B., Proynov, E. I., Pieniazek, P. A., Rhee, Y. M., Ritchie, J., Rosta, E., Sherrill, C. D., Simmonett, A. C., Subotnik, J. E., Woodcock, H. L., Zhang, W., Bell, A. T., Chakraborty, A. K., Chipman, D. M., Keil, F. J., Warshel, A., Hehre, W. J., Schaefer, H. F., Kong, J., Krylov, A. I., Gill, P. M., Head-Gordon, M. *Q-Chem*, version 3.0; Q-Chem, Inc.: Pittsburgh, PA, 2006.

(61) Ponder, J. W. *TINKER, Software Tools for Molecular Design*, version 4.2; 2004.

(62) Zhang, Y. *J. Chem. Phys.* **2005**, *122*, 024114.

(63) Zhang, Y. *Theor. Chem. Acc.* **2006**, *116*, 43–50.

(64) Zhang, Y.; Lee, T. S.; Yang, W. *J. Chem. Phys.* **1999**, *110*, 46–54.

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

(66) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308–313.

(67) Vannini, A.; Volpari, C.; Gallinari, P.; Jones, P.; Mattu, M.; Carfi, A.; De Francesco, R.; Steinkuhler, C.; Di Marco, S. *Embo Rep.* **2007**, *8*, 879–884.

(68) Schuetz, A.; Min, J.; Allali-Hassani, A.; Schapira, M.; Shuen, M.; Loppnau, P.; Mazitschek, R.; Kwiatkowski, N. P.; Lewis, T. A.; Maglathin, R. L.; McLean, T. H.; Bochkarev, A.; Plotnikov, A. N.; Vedadi, M.; Arrowsmith, C. H. *J. Biol. Chem.* **2008**, *283*, 11355–11363.

(69) Budayova-Spano, M.; Fisher, S. Z.; Dauvergne, M. T.; Agbandje-McKenna, M.; Silverman, D. N.; Myles, D. A. A.; McKenna, R. *Acta Crystallographica Section F-Structural Biology and Crystallization Communications* **2006**, *62*, 6–9.

(70) Yoshida, H.; Yamada, M.; Ohyama, Y.; Takada, G.; Izumori, K.; Kamitori, S. *J. Mol. Biol.* **2007**, *365*, 1505–1516.

(71) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(72) Wu, R.; Xie, H.; Mo, Y.; Cao, Z. *J. Phys. Chem. A* **2009**, *113*, 11595–11603.

CT100525R

JCTC Journal of Chemical Theory and Computation

Electronic Excitations of Simple Cyanine Dyes: Reconciling Density Functional and Wave Function Methods

Robert Send,*,† Omar Valsson,*,‡ and Claudia Filippi*,‡

Institut für Physikalische Chemie, Karlsruher Institut für Technologie, Kaiserstraβe 12, 76131 Karlsruhe, Germany, and Faculty of Science and Technology and MESA+ Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received November 3, 2010

Abstract: The simplest cyanine dye series $[H_2N(CH)_nNH_2]^+$ with $n = 1$, 3, 5, 7, and 9 appears to be a challenge for all theoretical excited-state methods since the experimental spectra are difficult to predict and the observed deviations cannot be easily explained with standard arguments. We compute here the lowest vertical excitation energies of these dyes using a variety of approaches, namely, complete active space second-order perturbation theory (CASPT2), quantum Monte Carlo methods (QMC), coupled cluster linear response up to third approximate order (CC3), and various flavors of time-dependent density functional theory (TDDFT), including the recently proposed perturbative correction scheme (B2PLYP). In our calculations, all parameters such as basis set, active space, and geometry dependence are carefully analyzed. We find that all wave function methods give reasonably close excitation energies, with CASPT2 yielding the lowest values, and that the B2PLYP scheme gives excitations in satisfactory agreement with CC3 and DMC, significantly improving on the generalized gradient and hybrid approximations. Finally, to resolve the remaining discrepancy between predicted excitation energies and experimental absorption spectra, we also investigate the effect of excited-state relaxation. Our results indicate that a direct comparison of the experimental absorption maxima and the theoretical vertical excitations is not possible due to the presence of nonvertical transitions. The apparent agreement of earlier CASPT2 calculations with experiments was an artifact of the choice of active space and the use of an older definition of the zero-order Hamiltonian.

1. Introduction

Cyanine dyes are characterized by a conjugated π-electron system connecting two nitrogen atoms and carrying a positive charge.[1] They are naturally occurring as red colorants in fly agaric mushrooms or red beets[2] and are of great industrial interest for their application in solar cells,[3] optical storage media (CDs, DVDs),[4] cancer cell recognition,[5] nonlinear optics,[6] and as biomarkers for nucleic acid detection.[7] This wide range of important applications has made cyanine dyes an early target of theoretical studies aimed at demonstrating the predictive power of computational approaches.[8]

In the past two decades, efficient computational approaches for excited states have been developed, which allow the description of large dyes and the fast screening of molecular libraries in search of specific excited-state properties.[9] In particular, time-dependent density functional theory (TDDFT)[10−12] has become the method of choice for the study of large molecular systems and has been successfully employed to search for highly specialized chromophores and investigate several dye

* E-mail: robert.send@kit.edu; o.valsson@utwente.nl; c.filippi@utwente.nl.

† Karlsruher Institut für Technologie.

‡ University of Twente.

10.1021/ct1006295
Published on Web 01/10/2011

families.[13,14] The efficiency of TDDFT comes in some cases at the price of lower accuracy as compared to conventional highly correlated quantum chemistry methods. It is, for instance, well-known that the description of excitations with charge-transfer, multireference, or Rydberg character is generally problematic in TDDFT. Since none of these features appears to characterize the lowest excited state of the cyanine dyes, one would expect TDDFT to be well suited for the description of this class of systems.

Surprisingly, as early as 2001, Schreiber et al.[15] showed that the excitation energies of the cyanine dyes obtained by TDDFT deviate by more than 1 eV from the values obtained with the CASPT2 method, which is often regarded as one of the most accurate excited-state approaches available. Since the examples chosen in ref 15 were the simplest models of cyanine dyes, the result suggests that TDDFT is not applicable to any member of this dye family. Until today, none of the available density functionals significantly improved the agreement with the reference CASPT2 values given in ref 15. The reasons for the large errors in the TDDFT results for the cyanine dyes are not understood.

Since the early work by Schreiber et al.,[15] excited-state methods have seen several important developments: (i) The efficient implementation of coupled-cluster (CC) response methods in combination with the resolution-of-the-identity (RI) approximation represents a powerful single-reference complement to TDDFT,[16] (ii) Efficient excited-state gradient methods render a large number of excited-state properties accessible.[17,18] (iii) Developments in algorithms and hardware allow for the use of larger basis sets and higher-level theories. (iv) Quantum Monte Carlo (QMC) methods can be used as an alternative to CASPT2 and independent validation of TDDFT.[19−22] (v) The CASPT2 method has been modified and generally improved by the introduction of a novel definition of the zeroth-order Hamiltonian.[23]

None of these developments have been fully exploited in recent calculations of the cyanine dyes, where most efforts have instead been directed to apply different flavors of density functionals in order to improve the excitations and gain insight into the shortcomings of TDDFT. Unfortunately, none of the used functionals has yielded significant improvement, and the insight gained has therefore been limited. The only exception is the B2PLYP scheme by Grimme, which incorporates a perturbative correction based on Kohn−Sham orbitals in a form similar to wave function treatments.[24,25] We note that the extensive excitation benchmark of wave function methods of ref 26 unfortunately does not include any member of the cyanine dye family.

The present work represents a comprehensive treatment of the simple cyanine dye series using several state-of-the-art excited-state methods such as CASPT2, QMC, and CC response methods up to third approximate order; TDDFT also in the long-range corrected and B2PLYP flavors; and the Tamm−Dancoff approximation. We give a detailed account of all parameters which may affect the calculation of the excitations in the various approaches. Our discussion focuses on the lowest bright excited state, and we enclose results for higher excited states in the Supporting Information.

All computational details are given in section 2. We describe the dependence of the excitation energies on the basis set and the method used to optimize the ground-state geometry in section 3.1. This is followed by the excitation energies calculated with CC methods (section 3.2), CASPT2 (section 3.3), QMC (section 3.4), and TDDFT (section 3.5). In section 4, we discuss the relative performance of the theoretical approaches and their comparison with experiments. Our conclusions are summarized in section 5.

2. Computational Details

The ground-state structures are optimized within Hartree−Fock (HF), second-order Møller−Plesset (MP2), and density functional theory (DFT). To compute the excitation energies, we employ coupled-cluster (CC) methods, time-dependent density functional theory (TDDFT), the complete active space self-consistent field (CASSCF) method with its perturbative extension (CASPT2), and quantum Monte Carlo (QMC) methods. The CC response calculations[27,28] are performed at the singles (CCS), singles and doubles (CCSD),[29] approximate second (CC2),[16,30−32] and approximate third (CC3)[33,34] orders. In the DFT calculations, the PBE,[35] PBE0,[36−38] CAM-B3LYP,[39] and B2PLYP[24,25] functionals are employed. The Tamm−Dancoff approximation is employed in some of the TDDFT calculations and denoted with the prefix TDA.[40]

The resolution-of-the-identity (RI) approximation[41] is used in all MP2 and in some CC2 calculations and is indicated by the abbreviations RI-MP2[42] and RI-CC2.[31] All RI-MP2, RI-CC, and DFT calculations are performed with the TURBOMOLE code.[43] B2PLYP calculations are based on an unreleased TURBOMOLE implementation and the additional on top program RICC by Grimme.[24,25] The CC and CAM-B3LYP excitation energies calculated without the RI approximation are obtained with the DALTON program suite.[44] The CAM-B3LYP excitation energy of the largest dye with the triple-ζ basis is computed with the Gaussian 09 code.[45]

The complete active space calculations are performed using MOLCAS 7.2.[46] In the CASPT2 calculations, we employ the default IPEA zero-order Hamiltonian[23] unless otherwise stated and indicate if an additional constant level shift[47] is added to the Hamiltonian. In the CASPT2 calculations, we do not correlate as many of the lowest σ orbitals, as there are heavy atoms in the molecule. For some models, we use the Cholesky decomposition of the two-electron integrals[48] with the threshold of 10^{-8}. The default convergence criteria are used for all calculations.

The program package CHAMP[49] is used for the QMC calculations. We employ scalar-relativistic energy-consistent Hartree−Fock pseudopotentials[50] where the carbon and nitrogen 1s electrons are replaced by a nonsingular s−nonlocal pseudopotential and the hydrogen potential is softened by removing the Coulomb divergence. Different Jastrow factors are used to describe the correlation with different atom types, and for each atom type, the Jastrow factor consists of an exponential of the sum of two fifth-order polynomials of the electron−nucleus and the electron−electron distances, respectively.[51] We also test the effect of including an

electron−electron−nuclear term. The starting determinantal components are obtained in CASSCF calculations, which are performed with the program GAMESS(US),[52] and the final CAS expansions are expressed on the CASSCF natural orbitals. The CAS wave functions of the states of interest may be truncated with an appropriate threshold on the CSF coefficients for use in the QMC calculations. The Jastrow correlation factor and the CI coefficients are optimized by energy minimization within VMC, and when indicated in the text, also the orbitals are optimized along with the Jastrow and CI parameters. The pseudopotentials are treated beyond the locality approximation,[53] and an imaginary time step of 0.05 au is used in the DMC calculations.

2.1. Basis Sets and Ground-State Structures. To investigate the basis-set dependence of the ground-state structures and of the CC and TDDFT excitations, we use the ANO-L-VXZP basis sets[54] and Dunning's correlation consistent cc-pVXZ and aug-cc-pVXZ basis sets.[55−58] For the ANO basis sets, the MOLCAS contraction scheme is employed, namely, ANO-L-VDZP [3s2p1d]/[2s1p], ANO-L-VTZP [4s3p2d1f]/[3s2p1d], and ANO-L-VQZP [5s4p3d2f]/[4s3p1d]. The ANO-L-VXZP series is used in the CASSCF and CASPT2 calculations.

In the QMC calculations, we use the Gaussian basis sets[50] specifically constructed for our pseudopotentials. In particular, we employ the cc-pVDZ basis, denoted by D, and the T′ and Q′ basis sets, which consist of the cc-pVDZ for hydrogen combined respectively with the cc-pVTZ and cc-pVQZ basis sets for the heavy atoms. The D+, T′+, and Q′+ basis sets are constructed by augmenting the corresponding basis with diffuse s, p, and d functions[59] on the heavy atoms. Basis functions with higher angular momentum than d are not included in the T′, T′+, Q′, and Q′+ basis sets.

Unless indicated otherwise, the CC, CASPT2, and TDDFT excitation energies are calculated with the ANO-L-VTZP basis set and the QMC excitations with the T′+ basis set. All excitation energies are computed on the RI-MP2/cc-pVQZ ground-state structures with the exception of the TDDFT excitations, which are obtained using the PBE0/cc-pVQZ structures.

2.2. Auxiliary Basis Sets. In the RI-MP2/ANO-L-VXZP and RI-CC2/ANO-L-VXZP calculations, the corresponding auxiliary basis sets are not available. To assess the impact of using the ANO-L-VXZP basis sets in combination with the available aug-cc-pVXZ auxiliary basis sets, we calculate the error in the correlation energy introduced by the RI approximation for carbon and nitrogen atoms and for H_2. The quantity commonly used to access the quality of an auxiliary basis set is defined as

$$\alpha = \frac{\delta_{\mathrm{RI}}}{|\Delta E(\mathrm{MP2})|} \tag{1}$$

where ΔE(MP2) is the MP2 correlation energy and δ_{RI} is given by

$$\delta_{\mathrm{RI}} = \frac{1}{4}\sum_{i<j}^{\mathrm{occ.}}\sum_{a<b}^{\mathrm{virt.}} \frac{\left|\langle ab||ij\rangle_{\mathrm{exact}} - \langle ab||ij\rangle_{\mathrm{RI}}\right|^2}{\varepsilon_a - \varepsilon_i + \varepsilon_b - \varepsilon_j} \tag{2}$$

The values of α obtained by combining the ANO-L-VXZP basis with the auxiliary aug-cc-pVXZ basis sets are given in the Supporting Information. When the aug-cc-pVQZ auxiliary basis is employed, $\alpha < 0.05$ ppm, which is in line with standard auxiliary-basis-optimization conditions.[58] Therefore, we adopt this auxiliary basis in all of our RI calculations.

2.3. Extrapolation of Excitation Energies. The extrapolated CC3/ANO-L-VTZP (exCC3) excitation energies are obtained as

$$\tilde{E}_{\mathrm{T}}^{\mathrm{CC3}} = E_{\mathrm{T}}^{\mathrm{CC2}} + (E_{\mathrm{D}}^{\mathrm{CC3}} - E_{\mathrm{D}}^{\mathrm{CC2}}) \tag{3}$$

This extrapolation formula is motivated by the observation that triple excitations are less basis-set-sensitive than single and double excitations.[60−62]

3. Vertical Excitation Energies

The cyanine dye molecules studied in this work are shown in Figure 1. We consider hydrogen-terminated dyes of increasing size, which we denote as CN3, CN5, CN7, CN9, and CN11. All hydrogen-terminated dyes have C_{2v} symmetry. For these molecules, we also construct the equivalent dyes where the terminating hydrogens are substituted by methyl groups.

3.1. Basis Set Convergence and Geometry Dependence. We employ the cc-pVXZ, aug-cc-pVXZ, and ANO-L-VXZP series to investigate the basis set dependence of the CC and CASPT2 excitations and give a complete survey of the results in the Supporting Information (SI). In this section, we focus on the smallest molecule, CN3, since it displays the largest dependence on the basis set. The basis-set dependence of the TDDFT and QMC excitations will be discussed separately.

The CC2 excitations of CN3 as a function of the basis set are shown in Figure 2. The correlation-consistent basis series gives the slowest convergence in the excitation energy as a function of basis set size, and an error which is still as large as 0.15 eV when a quadruple-ζ basis is employed. The inclusion of augmentation completely cures the problem since the energy obtained with the double-ζ basis only differs from the augmented quadruple-ζ value by 0.02 eV. The excitations computed with the ANO series converge similarly to the augmented correlation consistent values, and the use of a triple-ζ basis yields the quadruple-ζ value within better than 0.01 eV.

The behavior of the CASPT2 excitations as a function of the basis set is shown for CN3 in Figure 3. The excitations are obtained with the standard IPEA Hamiltonian (S-IPEA) as well as with the IPEA shift set to zero (0-IPEA), as in versions of MOLCAS prior to 6.4. The energies obtained with the IPEA Hamiltonian are 0.2 eV higher than the values obtained without the shift, and the difference is independent of the choice of the basis set. The behavior of the CASPT2 values with and without the IPEA shift closely parallels what is observed for the CC2 excitations. In particular, the inclusion of diffuse augmentation is absolutely necessary

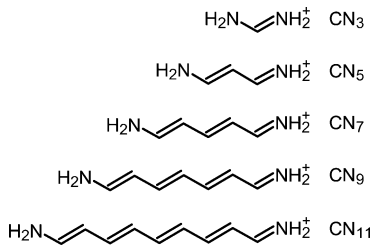


Figure 1. Hydrogen-terminated cyanine dyes considered in this work. Only one of the two resonant structures of each molecule is shown. The other structure can be obtained by having the first double bond at the other nitrogen atom.

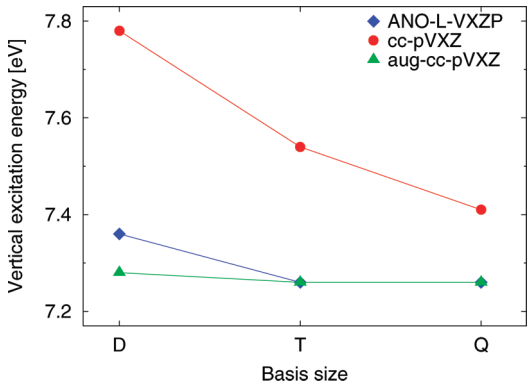


Figure 2. CC2 vertical excitation energies of CN3 computed with different basis sets. The ground-state MP2/cc-pVQZ geometry is used.

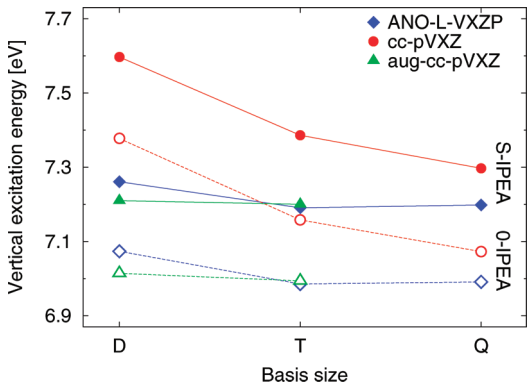


Figure 3. CASPT2 vertical excitation energies of CN3 computed with (S-IPEA) and without (0-IPEA) IPEA shift, and different basis sets. The ground-state MP2/cc-pVQZ geometry is used.

when employing the correlation consistent series, while the ANO energies are well converged when a triple-ζ basis is employed.

The optimal basis set for the present system is a correlation-consistent triple-ζ basis with diffuse augmentation or an ANO triple-ζ basis. Depending on the program, segmented or generally contracted basis sets can be more efficient. As MOLCAS is optimized for generally contracted basis sets, the discussion in the following is based on ANO triple-ζ basis sets. These give CC2 and CASPT2 excitations which are well converged in the basis sets for CN3 as well as for the other molecules (see the SI). In the SI, we also include excitation energies calculated with the correlation consistent Dunning basis sets, more common in CC calculations. The

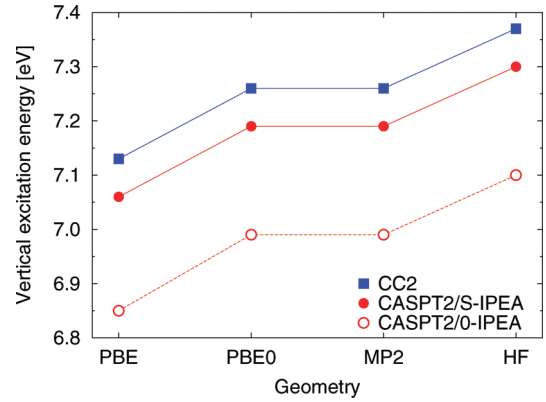


Figure 4. CC2 and CASPT2 vertical excitation energies of CN3 computed on different geometries. The ANO-L-VTZP basis is used.

most efficient choice in segmented contracted basis sets is the recent property-optimized basis sets by Rappoport and Furche.[63] These became available very recently, so we only include a table with the corresponding excitations in the SI.

The dependence of the CC2 and CASPT2 excitation energies on the method employed to optimize the ground-state geometry is shown for CN3 in Figure 4. As in the case of the basis set size, the dependence is most significant for the smallest molecule, CN3, as shown at the CC2 level in the SI. Indepedently of the approach used to compute the excitations and for all chain lengths, PBE and HF geometries give the lowest and highest excitations, respectively, while PBE0 and MP2 geometries are in between. The largest difference between the excitations computed on PBE and HF geometries is 0.24 eV at the CC2 level, as obtained for CN3, and is comparable at the CASPT2 level. The use of PBE0 and MP2 geometries gives very similar excitations with the largest difference of 0.03 eV obtained for CN5. Throughout this work, we use MP2 or PBE0 geometries to reduce the influence of the choice of the ground-state structure on the vertical excitation and focus on the performance of the approach employed to compute the excitations.

3.2. Coupled Cluster Results. For all dyes, we give the convergence of the CC excitation energies with respect to the size of the ANO basis set and the order of the CC expansion in Table 1. As already discussed in section 3.1, the triple-ζ basis set is the most cost efficient choice as an increase to quadruple-ζ only changes the excitation energies by less than 0.01 eV. The CC3 calculations for the largest dye, CN11, are not feasible at the ANO-L-VTZP level, so we also compute the triple-ζ extrapolated CC3 results (exCC3) using eq 3. When available, the CC3 results deviate from their extrapolated counterparts by less than 0.03 eV, and the error in the extrapolation is therefore comparable to the residual basis-set error.

The behavior of the excitation energies at different CC levels reflects the typical convergence of the correlation energy contribution.[26] With the ANO-L-VTZP basis, this convergence is characterized by an increase of less than 0.03 eV when going from CC2 to the full inclusion of doubles amplitudes in CCSD and a decrease of less than 0.14 eV when going from CC2 to CC3. The decrease in excitation energies when going from CC2 to CC3 is larger than the

Table 1. Coupled Cluster Vertical Excitation Energies (eV) for the 1^1B_1 State of the Cyanine Dye Series Computed at the CC2, CCSD, and CC3 Levels with the ANO-L-VXZP Basis Sets[a]

molecule	basis	CC2	CCSD	CC3	exCC3
CN3	D	7.36	7.32	7.27	
	T	7.26	7.29	7.18	7.16
	Q	7.26	7.30	7.18	
CN5	D	5.02	4.98	4.89	
	T	4.97	4.98	4.86	4.84
	Q	4.96	4.99	4.86	
CN7	D	3.83	3.79	3.69	
	T	3.79	3.81	3.68	3.65
CN9	D	3.13	3.09	2.99	
	T	3.10	3.11		2.96
CN11	D	2.66	2.62	2.52	
	T	2.64[b]			2.53

[a] The extrapolated CC3 values (exCC3) are obtained by adding the difference between the double- ζ CC3 and CC2 values to the triple- ζ CC2 results. The ground-state RI-MP2/cc-pVQZ structures are employed. [b] Computed with the RI approximation.

one observed for the corresponding bright state in butadiene (0.04 eV) or in the protonated Schiff base models (0.01 eV).[64,65] The T_1 diagnostic[66] remains lower than the empirical threshold of 0.02, indicating that the Hartree−Fock determinant is a good zeroth-order description of the ground state, and CC2 and CCSD results can therefore be considered reliable.

Further insight into our calculations can be gained by the amount of single- and double-excitation contribution in the CC3 excitation energies. The single-excitation contributions decrease from 89% to 84% when going from CN3 to CN11. The double-excitation contributions increase from 11% to 16% when going from CN3 to CN11. This finding is in line with the growing difference between CC2 and CC3 results upon lengthening of the chain. The correlation energy strongly depends on double excitations for all molecules, and triple excitations contribute more than in the analogous polyenes and protonated Schiff bases. The ground-state correlation energy shows, on the other hand, little dependence on the chain length. For all molecules, 92% of the CC3 correlation energy is obtained already at the CC2 level, and the CC3 correlation energy per electron is identical up to 0.1 mH for all dyes. This finding indicates that electron correlation effects are important mainly in the description of the excitation, for which an accurate description of correlation is therefore essential.

3.3. CASPT2 Results. The choice of the active space significantly affects the CASPT2 energies of the cyanine dyes, particularly of the smallest ones. As shown below, previous calculations[15] employed active spaces that were too small and led to underestimated CASPT2 excitation energies.

We extensively investigated the dependence of the excitations on the choice of the active space, and we give a complete account of our calculations in the SI. In Table 2, we present the most relevant subset of our results where the number of active π orbitals of a_2 and b_2 symmetry included in the CAS is l times the number of heavy atoms. This construction corresponds to l atomic orbitals of p character per heavy atom and produces a series of balanced active spaces. We observe that choosing l equal to 2 offers a good

Table 2. CASPT2 Vertical Excitations (eV) of the 1^1B_1 State Computed with (S-IPEA) and without (0-IPEA) IPEA Shift and with Different CAS(m,n) Expansions[a]

	CAS(m,n)			CASPT2	
molecule	m [a_2, b_2]	n [a_2, b_2]	CASSCF	0-IPEA	S-IPEA
CN3	4 [2, 2]	3 [1, 2]	8.12	6.55	6.90
		6 [2, 4]	7.56	6.99	7.19
		9 [3, 6]	7.63	6.97	7.14
CN5	6 [2, 4]	5 [2, 3]	5.46	4.23	4.62
		10 [4, 6]	5.32	4.46	4.69
		15 [6, 9]	5.33	4.49	4.68
CN7	8 [4, 4]	7 [3, 4]	3.92	3.17	3.56
		14 [6, 8]	3.91	3.30	3.52
		21 [9, 12]	3.96	3.30	3.49
CN9[b]	10 [4, 6]	9 [4, 5]	2.99	2.55	2.92
		18 [8, 10]	3.13	2.59	2.81
CN11[b]	12 [6, 6]	**11 [5, 6]**	2.39	2.10	2.46

[a] All π electrons (m) in the reference are included, and the number of active π orbitals is $n = i + j$, where i and j are orbitals of a_2 and b_2 symmetry, respectively, as specified by the notation [i, j]. The number of active orbitals is a multiple l of the number of heavy atoms as obtained by using l atomic orbitals of p character per heavy atom. We denote in boldface the optimal choice of active space in cost and accuracy for CN3−CN7. For CN9 and CN11, the maximum feasible values of l are 2 and 1, respectively. Additional active spaces not constructed as a multiple of l are listed in the SI. The ANO-L-VTZP basis set and the ground-state RI-MP2/cc-pVQZ structures are employed. [b] Cholesky decomposition with 10^{-8} threshold.

compromise between accuracy and computational cost since the corresponding excitations are always converged to better than 0.05 eV. For the largest dye, CN11, we cannot perform a calculation with l equal to 2, as the use of 22 active orbitals is not feasible. However, the excitation energy of CN11 computed with l equal to 1 is converged within 0.05 eV, as can be seen from the excitations computed with larger CAS dimensions given in the SI.

Our optimal active space must be contrasted to the use of an active space with an equal number of active electrons and active orbitals, as adopted in ref 15. We illustrate the shortcomings of this alternative construction by plotting the excitation of CN3 as a function of the dimension of the active space in Figure 5. The use of a CAS(4,4) space as in ref 15 yields an excitation which is underestimated by as much as 0.4 eV, while the excitation computed with a CAS(4,6) expansion is perfectly well converged. The dependence on the size of the CAS is slightly more pronounced when the zero-order Hamiltonian with no IPEA shift is employed as in ref 15, and as expected, the difference between the excitations computed with and without IPEA shift diminishes with increasing CAS size.

We summarize the CASPT2 excitations for our optimal choice of active space as a function of the ANO basis sets in Table 3. As discussed previously, the use of an ANO triple-ζ basis gives well converged excitations whether one uses the zero-order Hamiltonian with or without the IPEA shift. The excitations computed without the IPEA shift are 0.2 eV lower than the values obtained with the standard IPEA Hamiltonian, independent of the basis. For CN11, the difference between the values computed with and without IPEA shift appears to be larger than for the smaller dyes, and equal to 0.36 eV. The use of larger active spaces would

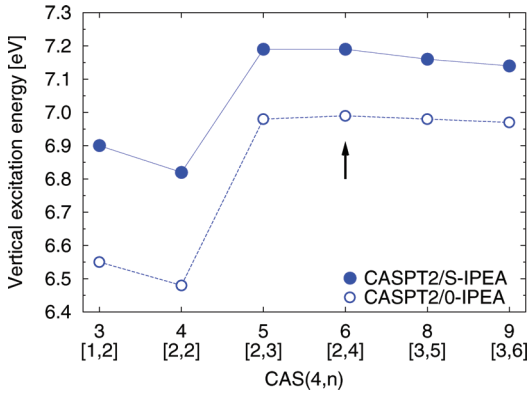


Figure 5. CASPT2 vertical excitation energies of CN3 computed with (S-IPEA) and without (0-IPEA) IPEA shift and with different CAS(4,*n*) expansions. The number of π electrons in the reference configuration is 4. The number of active orbitals is $n = i + j$, and i and j are orbitals of a_2 and b_2 symmetry, respectively, as specified by the label [*i*,*j*]. The arrow indicates a balanced CAS size, which corresponds to the use of two atomic orbitals of p character per heavy atom and represents an optimal compromise in accuracy and cost. The CAS(4,4) chosen in ref 15 is clearly inadequate. The ground-state MP2/cc-pVQZ geometry is used.

Table 3. CASSCF and CASPT2 Vertical Excitation Energies (eV) for the 1^1B_1 State of the Cyanine Dye Series Computed with the ANO-L-VXZP Basis Sets and the Optimal Active Space[a]

		CAS(*n*,*m*)			CASPT2	
molecule	basis	*n* [a_2, b_2]	*m* [a_2, b_2]	CASSCF	0-IPEA	S-IPEA
CN3	D	4 [2, 2]	6 [2, 4]	7.59	7.07	7.26
	T			7.56	6.99	7.19
	Q			7.56	6.99	7.20
CN5	D	6 [2, 4]	10 [4, 6]	5.25	4.53	4.74
	T			5.32	4.46	4.69
	Q			5.32	4.46	4.69
CN7	D	8 [4, 4]	14 [6, 8]	3.85	3.35	3.55
	T			3.91	3.30	3.52
	Q[b]			3.92	3.30	3.53
CN9	D[b]	10 [4, 6]	18 [8, 10]	3.08	2.63	2.83
	T[b]			3.13	2.59	2.81
	Q[b]			3.14	2.59	2.81
CN11	D[b]	12 [6, 6]	11 [5, 6]	2.37	2.13	2.46
	T[b]			2.39	2.10	2.46

[a] A CAS(*n*,*m*) expansion is used to compute the ground- (1^1A_1) and excited-state (1^1B_1) energies, where *n* and *m* denote the number of electrons and molecular orbitals, respectively. The ground-state RI-MP2/cc-pVQZ structures are employed. [b] Obtained with the Cholesky decomposition with 10^{-8} threshold.

however reduce the difference to less than 0.25 eV also for CN11 (see the SI). This finding reflects the fact that CASPT2 excitations computed with the IPEA shift converge faster to the values obtained with larger CAS dimensions.

3.4. QMC Results. In the determinantal component of the Jastrow−Slater wave functions, we choose the active space identified as optimal in the CASPT2 calculations and always optimize at least the Jastrow and linear coefficients in energy minimization within variational Monte Carlo. Other ingredients in the trial wave function may impact the excitation, such as the choice of basis set, the truncation threshold on the CAS expansion, the form of the Jastrow factor, and whether one optimizes also the orbitals in the determinantal component. We investigate the effect of changing these parameters in the wave function and summarize the results in Table 4.

Most tests are performed for the smallest dye, CN3, whose excitation appears to be most sensitive to the features of the wave function. We find that including electron−electron−nucleus terms in the Jastrow factor has little effect on the excitation of CN3. While the VMC excitation slightly increases, the DMC excitation is unchanged by the presence of these additional terms in the Jastrow factor. Therefore, given the higher computational cost of these three-body terms, we only include electron−electron and electron−nucleus correlations in the Jastrow factor for all other dyes. Concerning the basis, we find that the D+ basis leads to excitations which are clearly overestimated in VMC, while T′+ gives converged excitations when compared to the Q′+ values both in VMC and DMC. Even though the shortcomings of a D+ basis are more visible for CN3 than for CN5, we employ a T′+ as the default basis to compute the excitations of all dyes.

More critical for CN3 is the choice of the truncation threshold on the CAS expansion, especially if one does not reoptimize the orbitals. When only the linear coefficients are reoptimized in the presence of the Jastrow factor, the DMC excitation obtained with the full CAS expansion is 0.1 eV lower than the value computed with a threshold of 0.02. If the orbitals are reoptimized, the DMC excitations computed with a full CAS and a truncated expansion become 0.1 and 0.2 eV lower than the corresponding values obtained with CASSCF orbitals, and one recovers the same DMC value when employing the full or truncated CAS expansion. For CN5, the optimization of the orbitals does not significantly affect the excitations, and reducing the truncation threshold on the CAS expansion has a smaller effect on the excitation than for CN3. Therefore, for the larger dyes, we do not reoptimize the orbitals but only make sure we have convergence with respect to the number of configuration state functions included in the determinantal component. For all dyes, we collect the best available QMC results computed with a T′+ and a two-body Jastrow factor in Table 5.

3.5. TDDFT Results. The TDDFT excitations are computed with the PBE, PBE0, and long-range corrected CAM-B3LYP functionals. We also employ the PBE0 hybrid functional in the Tamm−Dancoff approximation (TDA-PBE0) as well as the hybrid functional with a perturbative correction as proposed in Grimme's non-self-consistent B2PLYP scheme.

The TDDFT results are listed in Table 6, where we report the values computed with the ANO triple-ζ basis, which are converged with respect to the basis set to better than 0.02 eV (see the SI). The difference between the PBE and PBE0 functionals is largest for the smallest CN3 dye, where the PBE excitation is 0.22 eV lower than the PBE0 result. With increasing chain length, the PBE and PBE0 excitations approach each other, only differing by 0.06 eV for CN9. For all dyes, the CAM-B3LYP results lie between the PBE and PBE0 results, with PBE giving the lowest excitation. The Tamm−Dancoff approximation and the B2PLYP scheme significantly change the excitation energies of the cyanine

Table 4. VMC and DMC Vertical Excitation Energies (eV) for the 1^1B_1 State of the Cyanine Dye Series[a]

	CAS(*m*,*n*)				CSF/det.				
molecule	*m* [a_2, b_2]	*n* [a_2, b_2]	basis	thr.	1^1A_1	1^1B_1	CASSCF	VMC	DMC
CN3	4 [2, 2]	6 [2, 4]	T′+	0.02	7/11	8/22	7.62	7.58(1)	7.58(2)
			T′+	0.02	7/11	8/22	7.62	7.63(1)	7.58(2)[b]
			T′+	0.02	7/11	8/22	7.62	7.47(1)	7.40(2)[c]
			D+	0.00	57/113	48/144	7.63	7.61(1)	7.50(2)
			T′+	0.00	57/113	48/144	7.62	7.52(1)	7.48(2)
			T′+	0.00	57/113	48/144	7.55	7.52(1)	7.48(2)[d]
			Q′+	0.00	57/113	48/144	7.58	7.51(1)	7.46(2)
			T′+	0.00	57/113	48/144	7.62	7.56(1)	7.47(2)[b]
			T′+	0.00	57/113	48/144	7.62	7.48(1)	7.38(2)[c]
CN5	6 [2, 4]	10 [4, 6]	T′+	0.08	4/7	5/12	5.30	5.21(1)	5.11(2)
			T′+	0.04	8/17	14/38	5.30	5.13(1)	5.05(2)
			D+	0.02	20/51	27/102	5.29	5.13(1)	5.08(2)
			T′+	0.02	22/59	28/106	5.30	5.15(1)	5.04(2)
			T′+	0.02	22/59	28/106	5.30	5.09(1)	5.03(2)[c]
CN7	8 [4, 4]	14 [6, 8]	T′+	0.02	40/111	42/156	3.89	3.90(1)	3.83(2)
CN9	10 [4, 6]	9 [4, 5]	T′+	0.04	13/39	17/42	2.98	3.22(1)	3.11(2)
			T′+	0.02	43/101	65/254	2.98	3.18(1)	3.09(2)
CN11	12 [6, 6]	11 [5, 6]	T′+	0.04	17/54	21/98	2.37	2.68(2)	2.62(2)

[a] A CAS(*m*,*n*) expansion is used to compute the ground-state (1^1A_1) and excited-state (1^1B_1) energies, where *m* and *n* denote the number of electrons and molecular orbitals, respectively. The threshold on the expansion and the corresponding number of CSFs and determinants are also listed. Unless indicated, only the Jastrow and CI parameters are optimized, and the Jastrow factor includes only electron−nuclear and electron−electron terms. The ground-state RI-MP2/cc-pVQZ structures are employed. [b] Including Jastrow e−e−n term. [c] Orbitals optimized including all external orbitals. [d] T′+ basis set with f functions.

Table 5. VMC and DMC Vertical Excitation Energies (eV) for the 1^1B_1 State of the Cyanine Dye Series[a]

	CAS(*m*,*n*)					
Molecule	*m* [a_2, b_2]	*n* [a_2, b_2]	CASSCF	VMC	DMC	
CN3	4 [2, 2]	6 [2, 4]	7.62	7.48(1)	7.38(2)	[b,c]
CN5	6 [2, 4]	10 [4, 6]	5.30	5.09(1)	5.03(2)	[b,d]
CN7	8 [4, 4]	14 [6, 8]	3.89	3.90(1)	3.83(2)	[d]
CN9	10 [4, 6]	9 [4, 5]	2.98	3.18(1)	3.09(2)	[d]
CN11	12 [6, 6]	11 [5, 6]	2.37	2.68(2)	2.62(2)	[e]

[a] For each molecule, we show the best available value from the QMC calculations obtained using the T′+ basis set and a Jastrow factor including electron−nuclear and electron−electron terms. A CAS(*m*,*n*) expansion is used to compute the ground-state (1^1A_1) and excited-state (1^1B_1) energies, where *m* and *n* denote the number of electrons and molecular orbitals, respectively. The threshold on the expansion is also listed. Unless indicated, only the Jastrow and CI parameters are optimized. The ground-state RI-MP2/cc-pVQZ structures are employed. [b] Orbitals optimized including all external orbitals. [c] Thr. of 0.0. [d] Thr. of 0.02. [e] Thr. of 0.04.

Table 6. TDDFT Excitation Energies (eV) of the Cyanine Dye Series Computed with the ANO-L-VTZP Basis Set and Different Functionals[a]

molecule	PBE	PBE0	CAM-B3LYP	B2PLYP	TDA-PBE0
CN3	7.40	7.62	7.55	7.30	8.03
CN5	5.22	5.33	5.26	5.05	5.84
CN7	4.11	4.18	4.12	3.92	4.71
CN9	3.44	3.50	3.44	3.25	4.02
CN11	2.98	3.03	2.97	2.80	3.54

[a] The ground-state PBE0/cc-pVQZ structures are employed.

dyes, as already pointed out in ref 25. The TDA-PBE0 excitations are higher than the PBE0 results by about 0.4 eV for CN3 and 0.5 eV for the other dyes. The B2PLYP excitation energies are 0.25−0.32 eV lower than the PBE0 results.

The excitation of the smallest dye, CN3, shows the strongest dependence on the choice of the functional and, in particular, on the amount of exact exchange included in the functional. While this finding appears to support the suggestion in ref 25 that the self-interaction error is significant for these dyes, we note that the inclusion of exact exchange yields the same excitations as conventional generalized gradient approximations (GGA) for the larger dyes. Therefore, as we discuss in section 4, the discrepancy between TDDFT and correlated methods observed also for the larger dyes cannot be simply attributed to self-interaction error.

Finally, our results follow the general trend observed for a larger set of functionals by Jacquemin et al.,[67] namely, that GGA excitation energies are lower than long-range corrected hybrid-GGA values, while hybrid GGAs give the largest excitation energies. We also note that our excitation energies deviate less than 0.08 eV from those of ref 67, and these small differences can be attributed to the use of different basis sets and ground-state structures.

4. Discussion

In this section, we first focus on the relative performance of the theoretical approaches employed to compute the vertical excitation energies of the cyanine dyes and then discuss their comparison with the available absorption spectra in solution.

4.1. Theoretical Comparison. In Table 7, we summarize our most representative theoretical results for the vertical excitation energies of the cyanine dyes, namely, the extrapolated CC3 excitation energies (exCC3), the CASPT2 values computed with the standard IPEA Hamiltonian (CASPT2/S-IPEA), and the TDDFT energies obtained with the PBE0 functional and the B2PLYP scheme, all computed with the ANO-L-VTZP basis. We also list the best available DMC excitations computed with the T′+ basis set. For an extensive comparison with CC2 or CCSD, CASPT2 with no IPEA shift, and other DFT functionals and the dependence on the

Table 7. Vertical Excitation Energies (eV) for the 1^1B_1 State of the Cyanine Dye Series[a]

Molecule	PBE0	B2PLYP	exCC3	CASPT2	DMC
CN3	7.62	7.30	7.16	7.19	7.38(2)
CN5	5.33	5.05	4.84	4.69	5.03(2)
CN7	4.18	3.92	3.65	3.52	3.83(2)
CN9	3.50	3.25	2.96	2.81	3.09(2)
CN11	3.03	2.80	2.53	2.46	2.62(2)

[a] The CC, CASPT2/S-IPEA, and TDDFT excitations are computed with the ANO-L-VTZP basis set. The best available QMC values obtained with the T′+ basis set are shown.

basis, CAS spaces, and geometries, we refer the reader to the previous sections.

Comparing wave function methods, CASPT2 gives the lowest and DMC the highest excitation energies, while exCC3 falls in between. This energetic order holds for all chain lengths except for CN3, where CASPT2 and exCC3 give almost identical results. The difference between CASPT2 and exCC3 ranges between 0.03 and 0.15 eV, and the differences are smallest for CN3 and CN11. The difference between DMC and exCC3 is of opposite sign and lies between 0.09 and 0.22 eV and decreases steadily from CN3 to CN11.

To establish the relative accuracy of the wave function approaches, we recall that the CASPT2 method is generally quite sensitive to the choice of the zero-order Hamiltonian. For the cyanine dyes, the use of a Hamiltonian with no IPEA shift, as was standard prior to MOLCAS 6.4, yields excitation energies that are on average 0.2 eV lower than the values obtained with the recommended IPEA shift of 0.25 (see Table 3). When the standard IPEA value is adopted, CASPT2 is in better agreement with other wave function methods, indicating that this novel definition of the zero-order Hamiltonian is more accurate and represents an improvement as compared to previous CASPT2 calculations.

Previous CASPT2 calculations of the cyanine dyes by Schreiber et al.[15] are also affected by another problem, namely, an inadequate choice of the CAS space (see Figure 5). The combined effect of the choice of zero-order Hamiltonian and the insufficient CAS dimension explains why the CASPT2 energies of ref 15 are underestimated, in particular for CN3, where their excitation of 6.63 eV must be compared to our value of 7.19 eV. Our excitations of the cyanine dye series should therefore be regarded as more reliable CASPT2 reference values due to the use of the IPEA Hamiltonian and a well converged size of active space.

The agreement between the exCC3 and DMC excitation energies is very satisfactory, with a difference of only 0.1 eV for the largest dyes. The larger discrepancy of 0.2 eV for the smallest dye can be explained by the fact that the high excitation of CN3 is clearly more sensitive to the description of static correlation and other parameters in the wave function. For CN3, the DMC calculations were performed employing the full active space and optimizing also the orbital parameters. The results are stable and further improvement not obvious. When comparing with DFT methods, we will refer to the exCC3 numbers, as they fall in between the CASPT2 and DMC, keeping in mind that the exCC3 excitations, in particular for the smallest dyes, might be slightly underestimated.

The TDDFT excitations computed with the hybrid GGA PBE0 are about 0.35−0.5 eV above the exCC3 results. As discussed in section 3.5, the use of the nonhybrid GGA PBE or the long-range corrected CAM-B3LYP does not lead to a significantly closer agreement with wave function methods. The same holds for the larger number of GGA functionals including the highly parametrized Minnesota functionals tested by Jacquemin et al.[67,68] These findings indicate that a closer agreement with wave function methods can only be obtained by going beyond the GGA and hybrid-GGA levels.

It is evident that the excitation energies of the cyanine dyes are sensitive to the correlation energy treatment. This can be seen in the spread observed among the wave function methods and, at the TDDFT level, from the TDA-PBE0 results. Application of the TDA further deteriorates the agreement with the wave function methods (see Table 6). Within the Tamm−Dancoff approximation, matrix elements that mix excitations and de-excitations are neglected so that the excited state is described by excitations only. It has already been stressed by Grimme and Neese[25] that the present cyanines are one of the rare cases where de-excitations substantially contribute to the excitation energy. Clearly, with the omission of the de-excitations, an important component of correlation energy is neglected.

The only TDDFT approach that significantly improves the agreement with the wave function methods is B2PLYP. The deviation from the exCC3 results ranges between 0.14 and 0.29 eV and increases from CN3 to CN11. The agreement between B2PLYP and DMC is almost perfect for the smaller dyes, and the difference increases to 0.2 for the larger models. Therefore, the discrepancy with either exCC3 or DMC increases for excitations that have a larger double excitation character (as seen in the exCC3 calculations). The improvement given by the use of the B2PLYP scheme comes however at the cost of an increase in computational scaling, the introduction of an additional empirical parameter, and other well-known limitations.[25]

The improved behavior of B2PLYP with respect to GGA or hybrid functionals can be understood from the presence of the additional perturbative correction. The non-self-consistent B2PLYP correction is analogous to the (D) correction in CIS(D) excitation energies, or the MP2 energy correction in the ground state, but computed with Kohn−Sham and not Hartree−Fock orbitals. B2PLYP is therefore an empirical perturbative way to incorporate some double-excitation character into the TDDFT excitation energies. In the ground state, the opposite-spin part of the MP2 energy correction is identical to the first nonvanishing order of the RPA correlation energy, as shown by Eshuis et al.[69] In the excited state, the good performance of B2PLYP is thus an indication that the use of exact RPA correlation may cure the shortcomings of TDDFT in the cyanine dyes by a satisfactory description of double-excitation character. A nonempirical route to incorporate double excitations into TDDFT has been formulated and applied for instance by Cave et al. on polyenes.[70]

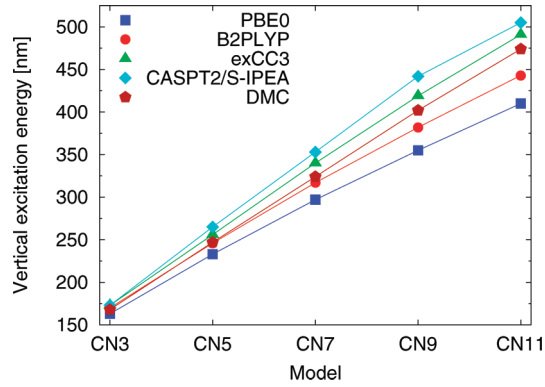


Figure 6. Vertical excitation energies of the cyanine dye series in nanometers. The exCC3, CASPT2/S-IPEA, and TDDFT results are computed with the ANO-L-VTZP basis set. The best available DMC values obtained with the T′+ basis set are shown.

Table 8. Experimental Absorption Maximum (eV) of the Cyanine Dye Series for Different Solutions and Substitutions at the Nitrogen Atoms (R_1, R_2)[a]

	nitrogen termination (R_1, R_2)		
molecule	(H, H)	(H, Me)	(Me, Me)
CN3			5.54[e]
CN5	4.34[c]	4.20,[c] 4.19[d]	3.97,[b] 4.01,[d] 3.96[e]
CN7	3.28[c]	3.15,[c] 3.14[d]	3.01,[f] 2.99,[b] 3.02,[d] 2.98[e]
CN9		2.53,[c] 2.51[d]	2.40,[b] 2.44,[d] 2.39[e]
CN11			1.96,[b] 2.03,[d] 1.98[e]

[a] The dielectric constant and the corresponding experimental temperature are given in brackets. [b] Ref 73, measured in methylendichloride (9.1, 20.0 °C). [c] Ref 74, measured in H_2O (80.4, 20.0 °C). [d] Ref 75, measured in methanol (32.6, 25 °C). [e] Ref 76, measured in methylendichloride (9.1, 20.0 °C). [f] Ref 77, measured in ethanol (24.3, 25 °C).

Finally, as is customary when discussing the cyanine dye series, we show the theoretical excitation energies in nanometers in Figure 6. All methods appear to follow the almost linear behavior traditionally called the vinyl shift.

4.2. Comparison with Experiments. We collect the experimental absorption maxima for comparison with the computed vertical excitation energies in Table 8. The experimental spectra were recorded in different solvents in the presence of ClO_4^- counterions, and all show broad absorption maxima for the lowest excited state. The position of the absorption maxima depends only negligibly on the dielectric constant of the solvent with variations smaller than 0.07 eV. Most experimental values were recorded for cyanine dyes with two methyl substituents on each nitrogen, and the absorption maxima of the methylated species are shifted to lower energies compared to the values of the unmethylated counterparts. The experimental methyl shift is 0.33−0.38 eV for CN5 and 0.26−0.30 eV for CN7 depending on the solvent.

As shown in Table 9, the experimental shifts upon methylation are theoretically well reproduced at the CC2 level with a value of 0.39 and 0.26 eV for CN5 and CN7, respectively, but largely overestimated by TDDFT/PBE0. The methyl shift in the CC2 excitations diminishes from 1.19 eV for CN3 to 0.17 eV for CN11, as expected since the influence of the end groups should vanish in large molecular

Table 9. RI-CC2 and TDDFT/PBE0 Excitation Energy (eV) of the 1^1B_1 State for the Methylated Streptocyanine Dye Series Computed with the ANO-L-VTZP Basis Sets[a]

	CC2		TDDFT/PBE0	
molecule	(H, H)	(Me, Me)	(H, H)	(Me, Me)
CN3	7.26	6.07	7.62	6.00
CN5	4.97	4.58	5.33	4.75
CN7	3.79	3.53	4.18	3.81
CN9	3.10	2.90	3.50	3.23
CN11	2.64	2.47	3.03	2.82

[a] The RI-MP2/cc-pVQZ and PBE0/cc-pVQZ ground-state structures in C_{2v} symmetry are employed for the CC2 and TDDFT calculations, respectively.

Table 10. Vertical and Constrained-Adiabatic Excitation Energies (eV) for the 1^1B_1 State, Obtained with RI-CC2 and the ANO-L-VTZP Basis Sets[a]

	E_{exc} (eV)		
molecule	vertical	adiabatic C_{2v}	Stokes shift C_{2v}
CN3	7.26	6.29	0.97
CN5	4.97	4.64	0.33
CN7	3.79	3.65	0.14
CN9	3.10	3.01	0.09
CN11	2.64	2.58	0.06

[a] Excited-state geometry optimizations are restricted to C_{2v} symmetry. The Stokes shift is the difference between vertical and C_{2v}-constrained adiabatic excitation energy. Relaxing the planarity constraint for CN3 and CN5 indicates that there is no planar minimum.

chains. The geometries of the methylated dyes are obtained in C_{2v} symmetry, but relaxing the symmetry constraint does not change the structure of the dyes, with the exception of CN3, where steric interaction between methyl groups at different nitrogens forces the CN3 dye into a nonplanar structure of C_2 symmetry. Recomputing the excitation energies at the CASPT2 and CC2 levels on the C_2 structure of CN3 only increases the excitations by 0.07 and 0.06 eV, respectively.

The basis for a comparison between computed vertical excitation energies and experimental absorption maxima is the assumption that the transition probability is largest at the ground-state minimum and when the transition is vertical, that is, when ground- and excited-state structures are identical. Examples where these assumptions are not satisfied are numerous,[71,72] but we restrict the discussion here to the validity of the comparison for the cyanine dye series.

The calculation of the absorption spectrum of the cyanine dyes using standard schemes is not possible here since it requires the existence of an excited-state minimum. Relaxing the excited state in C_{2v} symmetry at the CC2 level yields Stokes shifts of almost 1 eV, as shown in Table 10. Further relaxation of CN3 and CN5 without symmetry constraints leads to the highly twisted structures shown in Figure 7. Clearly, absorption spectra based on harmonic potentials in the excited state cannot be calculated for these systems.

The large Stokes shifts given in Table 10 explain the broad absorption maxima in the experiments. Within slight variation of the ground-state geometry, a large number of vibrational states at different energies can be reached if the Franck−Condon region of the excited state is distant from any

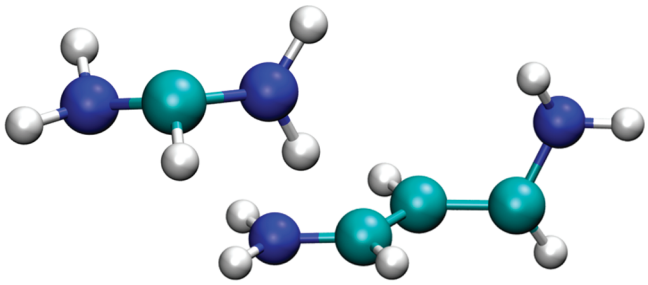


Figure 7. Excited-state minimal geometries of the CN3 and CN5 dyes obtained with RI-CC2 and the ANO-L-VTZP basis sets.

minimum. Moreover, the fact that a relaxed long-lived excited-state structure is most likely nonexistent also increases the likelihood of nonvertical transitions. We therefore conclude that the comparison between the computed vertical excitation energies and the experimental absorption maxima is not reliable, and certainly not suitable to assess the performance of high-level computational methods.

In fact, the direct comparison of calculated vertical excitation energies and experimental absorption maxima shows that CC2 results for the methylated dyes are on average 0.5 eV above the experimental data. The differences range from 0.44 eV for CN11 to 0.62 eV for CN5 and are dependent on the solvent. The deviations of the vertical excitations from the absorption maxima for the methylated species are consistent with the values obtained for the unmethylated dyes. Our CASPT2 excitation energies computed with the recommended IPEA zero-order Hamiltonian and carefully converged dimensions of the active space lie 0.34−0.35 eV above the experimental values. The apparently better agreement obtained in the older work by Schreiber et al. can be explained with their use of an inadequate active space as well as the use of a different zero-order Hamiltonian. The zero-order Hamiltonian used in our work was introduced a few years after the publication of Schreiber's results and is on average more accurate.

In summary, all methods give vertical excitation energies above the experimental absorption maxima with CASPT2 yielding the lowest values but still more than 0.3 eV higher than the experiments. The different wave function approaches yield very similar results, and all lie well above the experimental values. This supports our notion that the experimental absorption maxima correspond to nonvertical transitions. The influence of the solvent and the counterion are not included in our computational description and may further contribute to the discrepancy between theory and experiment.

5. Conclusion

For almost a decade, the simple cyanine dyes studied in this work have represented an intriguing and problematic case for TDDFT and a challenge for the development of new density functionals. The availability of accurate theoretical vertical excitations for these dyes is therefore very important in the assessment of the performance of existing or novel TDDFT approaches. With the present work, we offer carefully benchmarked reference values computed with CASPT2, QMC, CC, and various flavors of TDDFT as an aid for future developments. Our analysis based on such a large variety of excited-state methods gives a broad perspective on the parameters influencing the excited-state description.

We find that previous CASPT2 calculations[15] do not offer a reliable benchmark for the cyanine dyes since the chosen active space was inadequate and led to a severe underestimation of the CASPT2 excitations, with errors as large as 0.6 eV for the smallest CN3 dye. Our CASPT2 calculations are superior to these older studies in the use of the improved zero-order IPEA Hamiltonian and a balanced and well converged choice of the active space. Even though the empiricism introduced by the choice of zeroth-order Hamiltonian renders the assessment of CASPT2 calculations more difficult, the CASPT2 excitations obtained with the recommended IPEA shift appear to be more reliable than those computed without this shift, as the IPEA values are energetically closer to the extrapolated CC3 and DMC results. With our improved CASPT2 vertical excitations, we find that the agreement among all wave function methods is generally quite reasonable, with the largest deviations being observed for the smaller dyes, which appear most sensitive to the treatment of static correlation.

Consequently, the overestimation attributed in the past to TDDFT when comparing to older CASPT2 calculations is now not as severe. Nevertheless, the performance of standard GGA and hybrid GGA functionals is not satisfactory, and our calculations indicate that the discrepancy between TDDFT and wave function methods is due to an insufficient description of double-excitation character at the GGA level. The B2PLYP functional is an empirical scheme to partially incorporate double excitation character, and it significantly improves the description in the cyanine dyes, showing the best agreement with CC3 and DMC results.

Since all wave function methods are in close agreement and the calculations appear rather robust, we consider the corresponding excitations trustworthy. It therefore remains an open question as to why the theoretical results disagree with the location of the absorption maxima in the experimental spectra in solution. Quite surprisingly, we find that the addition of methylation significantly lowers the vertical excitations of the smallest dyes, bringing them in closer agreement with the experimental absorption maxima of the methylated species. Nevertheless, the remaining discrepancy between theory and experiment is quite large, and we attribute it to the presence of nonvertical transitions. In principle, one could prove or disprove this statement by a direct simulation of the absorption spectra. However, these simulations are not straightforward due to the lack of excited-state harmonic potentials and excited-state minima. We find that the relaxation of some of the smaller dyes in planar symmetry leads to Stokes shifts as large as 1 eV, and further unconstrained relaxation yields highly distorted structures that render the reconstruction of the spectra impossible. Clearly, a direct comparison of the experimental absorption maxima and the vertical excitation energies is not reliable and should not constitute the basis for the assessment of theoretical methods.

Acknowledgment. O.V. and C.F. acknowledge the support from the Stichting Nationale Computerfaciliteiten (NCF-NWO) for the use of the SARA supercomputer facilities. The work of R.S. was supported by the Center for Functional Nanostructures (CFN) of the Deutsche Forschungsgemeinschaft (DFG) within project C3.9. We thank Stefan Grimme for his TURBOMOLE implementation of the B2PLYP functional. R.S. acknowledges in-depth discussions with Henk Eshuis; helpful advise from Florian Weigend, Stephan Bernadotte, and Mikael P. Johansson; and critical comments on the manuscript by Filipp Furche.

Supporting Information Available: Assessment of auxiliary basis sets for the RI approximation. Oscillator strengths obtained with RI-CC2. Dependence of the vertical excitation energies on the choice of basis sets, geometry, and other relevant parameters in the different methods (e.g., active space or DFT functional). Dependence of CC2 and CASPT2 vertical excitations of the methylated dyes on the symmetry constraints. Ground-state structures and their dependence on the optimization method and basis set. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) *Light Absorption of Organic Colorants*; Fabian, J., Hartmann, H., Eds.; Springer: Heidelberg, 1980; Vol. 12, p 162.

(2) Strack, D.; Vogt, T.; Schliemann, W. *Phytochemistry* **2003**, *62*, 247.

(3) Mishra, A.; Fischer, M. K. R.; Bäuerle, P. *Angew. Chem., Int. Ed.* **2009**, *48*, 2474.

(4) Mustroph, H.; Stollenwerk, M.; Bressau, V. *Angew. Chem., Int. Ed.* **2006**, *45*, 2016.

(5) Weissleder, R.; Ntziachristos, V. *Nat. Med.* **2003**, *9*, 123.

(6) Würthner, F.; Schmidt, J.; Stolte, M.; Wortmann, R. *Angew. Chem.* **2006**, *118*, 3926.

(7) Ikeda, S.; Kubota, T.; Yuki, M.; Okamoto, A. *Angew. Chem., Int. Ed.* **2009**, *48*, 6480.

(8) Bayliss, N. S. *Q. Rev. Chem. Soc.* **1952**, *6*, 319.

(9) Paci, I.; Johnson, J. C.; Chen, X.; Rana, G.; Popovic, D.; David, D. E.; Nozik, A. J.; Ratner, M. A.; Michl, J. *J. Am. Chem. Soc.* **2006**, *128*, 16546.

(10) Casida, M. E. Time-dependent density-functional response theory for molecules. In *Recent Advances in Density Functional Methods, Part I*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p 155.

(11) Furche, F. *J. Chem. Phys.* **2001**, *114*, 5982.

(12) Rappoport, D.; Furche, F. Excited states and Photochemistry. In *Time-Dependent Density Functional Theory*; Marques, M. A. L., Ullrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K., Eds.; Springer: Berlin, 2006; Lecture Notes in Physics 706, p 337.

(13) Rappoport, D.; Furche, F. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6353.

(14) Fabian, J. *Dyes Pigm.* **2010**, *84*, 36.

(15) Schreiber, M.; Buβ, V.; Fülscher, M. P. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3906.

(16) Hättig, C.; Köhn, A. *J. Chem. Phys.* **2002**, *117*, 6939.

(17) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2007**, *126*, 201104.

(18) Send, R.; Furche, F. *J. Chem. Phys.* **2010**, *131*, 044107.

(19) Cordova, F.; Doriol, L. J.; Ipatov, A.; Casida, M. E.; Filippi, C.; Vela, A. *J. Chem. Phys.* **2007**, *127*, 164111.

(20) Tapavicza, E.; Tavernelli, I.; Rothlisberger, U.; Filippi, C.; Casida, M. E. *J. Chem. Phys.* **2008**, *129*, 124108.

(21) Filippi, C.; Zaccheddu, M.; Buda, F. *J. Chem. Theory Comput.* **2009**, *5*, 2074.

(22) Valsson, O.; Filippi, C. *J. Chem. Theory Comput.* **2010**, *6*, 1275.

(23) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142.

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

(25) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.

(26) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(27) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.

(28) Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333.

(29) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

(30) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.

(31) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.

(32) Hättig, C. *Adv. Quantum Chem.* **2005**, *50*, 37.

(33) Christiansen, O.; Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1995**, *103*, 7429.

(34) Koch, H.; Christiansen, O.; Jørgensen, P.; de Merás, A. M. S.; Helgaker, T. *J. Chem. Phys.* **1997**, *106*, 1808.

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

(36) Perdew, J. P.; Burke, K.; Ernzerhof, M. *J. Chem. Phys.* **1996**, *105*, 9982.

(37) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(38) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.

(39) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.

(40) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *314*, 291.

(41) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.

(42) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.

(43) *TURBOMOLE 6.2*, TURBOMOLE GmbH: Karlsruhe, Germany, 2010. http://www.turbomole.com (accessed Dec 2010).

(44) DALTON, an ab initio electronic structure program, release 2.0. See http://www.kjemi.uio.no/software/dalton/dalton.html (accessed Dec 2010).

(45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers,

E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J. ; Raghavachari, K. ; Rendell, A. ; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian Inc.: Wallingford, CT, 2009.

(46) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.

(47) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215.

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

(49) CHAMP is a quantum Monte Carlo program package written by Umrigar, C. J.; Filippi, C. and collaborators.

(50) Burkatzki, M.; Filippi, C.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 234105.

(51) Filippi, C.; Umrigar, C. J. *J. Chem. Phys.* **1996**, *105*, 213. As a Jastrow correlation factor, we use the exponential of the sum of three fifth-order polynomials of the electron−nuclear (e−n), the electron−electron (e−e), and the pure three-body mixed e−e and e−n distances, respectively. The Jastrow factor is adapted to deal with pseudoatoms, and the scaling factor κ is set to 0.6 au.

(52) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M. J. A. M., Jr. *J. Comput. Chem.* **1993**, *14*, 1347.

(53) Casula, M. *Phys. Rev. B* **2006**, *74*, 161102.

(54) Widmark, P.; Malmqvist, P.; Roos, B. O. *Theor. Chem. Acc.* **1990**, *77*, 291.

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

(56) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410.

(57) Wilson, A.; van Mourik, T.; Dunning, T. H., Jr. *THEOCHEM* **1997**, *388*, 339.

(58) Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057.

(59) We take the diffuse functions for the carbon and nitrogen atoms from the aug-cc-pVXZ basis sets in the EMSL Basis Set Library (http://bse.pnl.gov). In the D+ basis, we add the diffuse s and p functions from the aug-cc-pVDZ basis. In the T′+ and Q′+ basis sets, we add the diffuse s, p, and d functions from the aug-cc-pVTZ and aug-cc-pVQZ bases, respectively.

(60) Klopper, W.; Noga, J.; Koch, H.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 164.

(61) Karton, A.; Taylor, P. R.; Marian, J. M. L. *J. Chem. Phys.* **2007**, *127*, 064104.

(62) Johansson, M. P.; Olsen, J. *J. Chem. Theory Comput.* **2008**, *4*, 1460.

(63) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.

(64) Lehtonen, O.; Sundholm, D.; Send, R.; Johansson, M. P. *J. Chem. Phys.* **2009**, *131*, 024301.

(65) Send, R.; Sundholm, D.; Johansson, M. P.; Pawłowski, F. *J. Chem. Theory Comput.* **2009**, *5*, 2401.

(66) Lee, T. J.; Taylor, P. R. *Int. J. Quantum. Chem. Symp.* **1989**, *S23*, 199.

(67) Jacquemin, D.; Perpete, E. A.; Ciofini, I.; Adamo, C.; Valero, R.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2071.

(68) Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.

(69) Eshuis, H.; Yarkony, J.; Furche, F. *J. Chem. Phys.* **2010**, *132*, 234114.

(70) Cave, R. J.; Zhang, F.; Maitra, N. T.; Burke, K. *Chem. Phys. Lett.* **2004**, *389*, 39.

(71) Herzberg, G. *Molecular Spectra and Molecular Structure*; Van Nostrand and Reinhold,:New York, 1966, Vol. III, p 173.

(72) Dierksen, M.; Grimme, S. *J. Chem. Phys.* **2004**, *120*, 3544.

(73) Nikolajewski, H. E.; Dähne, S.; Leupold, D.; Hirsch, B. *Tetrahedron* **1968**, *24*, 6685.

(74) Grimm, B.; Dähne, S.; Bach, G. *J. Prakt. Chem.* **1975**, *317*, 161.

(75) Gürtler, O.; Dähne, S. *Z. Phys. Chem., Leipzig* **1974**, *255*, 501.

(76) Malhotra, S. S.; Whiting, M. C. *J. Chem. Soc.* **1960**, 3812.

(77) König, W.; Regner, W. *Ber. Dtsch. Chem. Ges.* **1930**, *63*, 2823.

CT1006295

JCTC Journal of Chemical Theory and Computation

Oscillator Strength: How Does TDDFT Compare to EOM-CCSD?

Marco Caricato,*,† Gary W. Trucks,† Michael J. Frisch,† and Kenneth B. Wiberg‡

Gaussian, Inc., 340 Quinnipiac St., Bldg. 40, Wallingford, Connecticut 06492, United States, and Department of Chemistry, Yale University, 225 Prospect St., New Haven, Connecticut 06511, United States

Received November 16, 2010

Abstract: In this work, we compare a large variety of density functionals against the equation of motion coupled cluster singles and doubles (EOM-CCSD) method for the calculation of oscillator strengths. Valence and Rydberg states are considered for a test set composed of 11 small organic molecules. In our previous work, the same systems and methods were tested against experimental results for the excitation energies. The results from this investigation confirm our previous findings, i.e., that there is a large difference between the functionals. For the oscillator strength, the average best agreement with EOM-CCSD is provided by CAM-B3LYP followed by LC-ωPBE and, to a lesser extent, B3P86 and LC-BLYP.

1. Introduction

Molecular UV/vis spectroscopy is routinely used in many areas of experimental research, and computational simulations of spectra have become an increasingly important, often essential tool for the interpretation of experiments. Nonetheless, the accurate simulation of molecular UV/vis spectra still represents a theoretically difficult challenge. Many methods and approximations have been developed to tackle this challenge. In particular, the advent of the time-dependent density functional theory (TDDFT)[1–3] within the adiabatic approximation has allowed the study of a large variety of molecules in different fields. Since the exact functional is not known, a myriad of approximate functionals has been proposed, and new ones are introduced every year. This constitutes a problem for the investigators, especially for nonspecialists, who want to take advantage of the computational efficacy of TDDFT.

Several papers have appeared in recent years that compare different functionals on different molecular systems for the calculation of vertical excitation energies.[4–16] In our previous work on this property,[12] we examined 11 small organic molecules for which extensive experimental and theoretical data in the gas phase are available:[17–29] a total of 69 states, 30 valence and 39 Rydberg in nature. Our set included a variety of density functionals: local spin density approximation (LSDA), generalized gradient approximation (GGA), GGA with kinetic energy density or meta-GGA (M-GGA), hybrid GGA (H-GGA), and hybrid meta-GGA (HM-GGA) as well as functionals that separate short- and long-range exchange contributions (with and without the correct long-range limits). Additionally, we considered four single reference wave function (WF) methods: configuration interaction with single excitations (CIS),[30] CIS with perturbative double excitations correction (CIS(D)),[31] random phase approximation (RPA),[32] and the highly accurate and computationally demanding equation of motion coupled cluster singles and doubles (EOM-CCSD).[33–35] Only single reference methods were considered because their results are unambiguous; therefore, they represent useful computational tools even for nontheoretically trained investigators. The computed data were compared to experimental results. As expected, EOM-CCSD was revealed to be the most accurate among the selected methods. This level of theory is often regarded as the best compromise between accuracy and computational cost for small- and medium-sized molecules, and it has the advantage that the quality of the wave function can be systematically improved by including more excitations in the wave operator.[33–35] On the other hand, there is not a systematic way to improve a particular functional, and its

* To whom correspondence should be addressed. E-mail: marco@gaussian.com.

† Gaussian, Inc.

‡ Yale University.

10.1021/ct100662n
Published on Web 12/22/2010

Table 1. List of Functionals Used in This Work

	year	type	% HF		year	type	% HF
LSDA[44,48]	1951	LSDA		B3VP86[45,47,48,50]	1993	H-GGA	20
BLYP[45,46]	1988	GGA		PBE1PBE[53,54,69,70]	1997	H-GGA	25
OLYP[49,46]	2001	GGA		THCTHHYB[57]	2002	HM-GGA	15
BP86[45,50]	1988	GGA		TPSSh[59,73]	2003	HM-GGA	10
BVP86[45,48,50]	1988	GGA		M05[60]	2005	HM-GGA	28
PBEPBE[53,54]	1997	GGA		BH&H,[44,46,48 a]	1993	H-GGA	50
HCTH[51,55,56]	2001	GGA		BH&HLYP,[44–46,48 a]	1993	H-GGA	50
THCTH[57]	2002	M-GGA		BMK[62]	2004	HM-GGA	42
VSXC[52]	1998	M-GGA		M05−2X[61]	2006	HM-GGA	56
TPSSTPSS[59]	2003	M-GGA		HSE1PBE[63]	2003	H-GGA	25−0[a]
O3LYP[46,49,58]	2001	H-GGA	11.61	CAM-B3LYP[64]	2004	H-GGA	19−65[b]
B3LYP[45–47,71,72]	1994	H-GGA	20	LC-BLYP[45,46,65,66]	2001	H-GGA	LC[c]
B3P86[45,47,50]	1993	H-GGA	20	LC-ωPBE[65–68]	2006	H-GGA	LC[c]

[a] Note that these are not the same as the half-and-half functionals proposed by Becke.[74] [b] Short-range−long-range. [c] The percentage of HF exchange increases as described in refs 65–68.

ability to compute a particular property must be tested against experimental results or a high ab initio level of theory. The results of our previous work[12] demonstrated that there are large differences between the functionals and that new functionals do not necessarily outperform older ones, at least for the range of transitions we studied. In fact, a relatively old functional, B3P86, performed as well as more recent ones such as CAM-B3LYP and LC-ωPBE; although the latter two were designed for the description of excited states, while the former was not.

Nevertheless, the excitation energy is only one part of a UV/vis spectrum; the other part is the intensity of the bands. This depends on the oscillator strength, which is related to the probability of the transition from the ground to an excited state when the molecule interacts with an electric field.[32] In the present literature, there are few comparisons of the performance of functionals for the calculation of this property.[7–10,36–43] In principle, the oscillator strength can be directly extracted from experimental results. Unfortunately, this is often difficult because of line broadening and overlapping of the excitation bands, and often the comparison with experimental oscillator strengths can only be qualitative. Therefore, in this work, the functionals are tested against the EOM-CCSD results. We only focus on single reference methods since their use is straightforward, as mentioned above. Although it is not expected that this level of theory is exact, EOM-CCSD is a well-defined reference[7–9] and can be systematically improved. The same states and molecules used in ref 12 are tested in the first part of the paper, and in the second part, a portion of the spectra is simulated by superimposing Gaussian line shapes in order to mimic the experimental bands (neglecting the vibronic structure, which is beyond the scope of the present work).

This paper is organized as follows. Section 2 describes the methods, the test systems, and other details of the calculations. Section 3 reports the results for the states used in ref 12, while section 4 contains the simulated spectra. An overall discussion of the results and concluding remarks are reported in section 5.

2. Computational Details

The selected methods are the same as those employed in ref 12. There are four WF methods: CIS, CIS(D), RPA, and EOM-CCSD. The approximate density functionals are listed in Table 1. The molecules we consider are also those from ref 12. There are three alkenes: ethylene (D_{2h}), isobutene (C_{2v}), and trans-1,3-butadiene (C_{2h}, we refer to this molecule simply as "butadiene" in the following); three carbonyl compounds: formaldehyde (C_{2v}), acetaldehyde (C_s), and acetone (C_{2v}); and five azabenzenes: pyridine (C_{2v}), pyrazine (D_{2h}), pyrimidine (C_{2v}), pyridazine (C_{2v}), and 1,2,4,5-tetrazine (D_{2h}, s-tetrazine). Their geometries were optimized at the MP2/6-311+G(d,p) level of theory (the geometries are available in the Supporting Information of ref 12). The vertical excitation energies and oscillator strengths are computed with the 6-311(3+,3+)G(d,p) basis set (the extra diffuse functions are available in the Supporting Information). All of the calculations are performed with a development version of the Gaussian suite of programs.[75] CIS(D) only corrects the CIS transition energy, but not the CIS transition dipole; therefore, the oscillator strengths for this method are computed as

$$f_i^{\mathrm{CIS(D)}} = \frac{2}{3}|\mu_{i0}^{\mathrm{CIS}}|^2 \Delta E_i^{\mathrm{CIS(D)}} \quad (1)$$

where f_i is the oscillator strength for the ith state, μ_{i0} is the transition dipole, and ΔE_i is the transition energy in atomic units.

The data in section 3 only refer to the states that we considered in ref 12, which have experimental data for the excitation energies. As outlined in section 1, experimental oscillator strengths are often nonquantitative; thus, the benchmark in this work is EOM-CCSD/6-311(3+,3+)G(d,p), which has shown great reliability in the study of excited states of small organic molecules.[7–9,76,77] We did not consider the transition properties computed with the linear response CCSD approach (LR-CCSD)[78,79] because it has been shown that the difference between LR- and EOM-CCSD is negligible for small molecules.[80,81] Since many states and many methods are compared, and oscillator strengths may differ by several orders of magnitude between different states, we perform a linear least-squares fit between the oscillator strengths computed with a particular method as a function of the reference (EOM-CCSD). The data reported in section 3 (X^L) are given as the slope of the line (with zero intercept)

that best fits the oscillator strengths computed with a particular method compared to EOM-CCSD:

$$f_i^{\text{method}} = X^L f_i^{\text{EOM-CCSD}} \quad (2)$$

We also report the R^2 parameter, which indicates how close the data points are to a line. This choice for the representation of the results provides the proper weight to the excited states according to the magnitude of their oscillator strengths. Indeed, many states in ref 12 have small oscillator strength or are symmetry-forbidden (they appear in the experimental spectrum because of vibrational symmetry breaking). All of the oscillator strengths are available in the Supporting Information.

Section 4 contains the simulated spectra obtained from the oscillator strengths by adding Gaussian line shapes through the Harada−Nakanishi equation:[82]

$$\varepsilon(\nu) = \frac{f_i}{3.483 \times 10^{-5}\sqrt{\pi}\sigma} \, e^{-((\nu - \nu_i)/\sigma)^2} \quad (3)$$

where ε is the extinction coefficient, ν is the excitation energy in eV, and σ is a parameter that we choose equal to 0.4 eV. We consider the region of the spectrum spanned by the first 20 states of EOM-CCSD ($\Delta E < 7-9$ eV, depending on the molecule). For the other methods, we include all the states necessary to have a complete description of the corresponding EOM-CCSD bands. Thus, the number of states considered is larger than in ref 12 and section 3. Furthermore, since including all of the methods would make the spectra unreadable, we only select 11 methods: CIS, LSDA, BLYP, B3LYP, B3P86, PBE1PBE, M05, CAM-B3LYP, LC-BLYP, LC-ωPBE, and EOM-CCSD. CIS and LSDA are the simplest among the WF and DFT methods. The other functionals are chosen among the various classes, in particular, those for which a long-range correction is available. We choose not to compare the simulated spectra with the experimental UV/vis spectra because the latter contain considerable vibronic structure, which strongly influences the shape and size of the bands. In contrast, the simple band structure provided by eq 3 would make the comparison with the experimental spectra difficult. Therefore, EOM-CCSD is the reference method also in this case.

3. Statistical Analysis

The values of X^L for alkenes, carbonyls, and azabenzenes are reported in Tables 2−4, respectively. These data are also grouped together in a graphical form in Figures 1 and 2. Figure 1 reports $X^L - 1$ for the single molecules, while Figure 2 reports $X^L - 1$ collectively for the groups of compounds (in Figure 1, the WF methods are not included because their X^L values are very large in many instances). In these figures, a negative value corresponds to a scaling factor smaller than 1, and a positive value to a factor larger than 1. The R^2 parameters are reported in Figures 3 and 4 for the single molecules and for the molecular groups, respectively (the numerical values are available in the Supporting Information). In the following discussion, we consider "good" the performance of a method that provides $0.66 < X^L < 1.5$ and $0.8 < R^2 \leq 1$. Obviously, the best performances are those where the X^L and R^2 parameters approach 1.

Table 2. X^L for the Alkenes

	ethylene	isobutene	butadiene	all
RPA	1.16	1.94	1.24	1.25
CIS	1.43	1.69	1.45	1.45
CIS(D)	1.49	1.65	1.47	1.48
LSDA	0.73	1.01	0.89	0.86
BLYP	0.53	0.21	0.77	0.69
OLYP	0.42	0.42	0.67	0.60
BP86	0.61	0.31	0.85	0.77
BVP86	0.61	0.31	0.85	0.77
PBEPBE	0.61	0.41	0.83	0.76
HCTH	0.61	0.71	0.85	0.79
THCTH	0.61	0.61	0.87	0.79
VSXC	0.72	0.13	0.90	0.83
TPSSTPSS	0.62	0.28	0.84	0.76
O3LYP	0.48	0.57	0.80	0.71
B3LYP	0.75	0.78	0.92	0.87
B3P86	0.88	1.35	0.97	0.97
B3VP86	0.79	0.81	0.95	0.91
PBE1PBE	0.81	0.93	0.96	0.92
THCTHHYB	0.72	0.78	0.93	0.87
TPSSh	0.55	0.50	0.90	0.80
M05	0.85	1.25	0.95	0.94
BH&H	0.96	1.38	1.07	1.05
BH&HLYP	0.96	1.37	1.07	1.05
BMK	0.88	1.05	1.01	0.98
M05−2X	1.00	2.11	1.06	1.09
HSE1PBE	0.83	0.99	0.96	0.93
CAM-B3LYP	0.95	1.38	1.02	1.02
LC-BLYP	1.06	2.15	1.12	1.15
LC-ωPBE	1.03	2.10	1.09	1.12

Table 3. X^L for the Carbonyls

	formaldehyde	acetaldehyde	acetone	all
RPA	3.01	1.80	3.32	2.53
CIS	2.72	1.35	5.09	2.65
CIS(D)	2.25	1.01	4.30	2.16
LSDA	0.43	0.74	0.25	0.53
BLYP	0.26	0.50	0.26	0.37
OLYP	0.43	0.62	0.26	0.48
BP86	0.23	0.43	0.26	0.33
BVP86	0.23	0.43	0.26	0.33
PBEPBE	0.24	0.55	0.26	0.38
HCTH	0.26	0.78	0.29	0.50
THCTH	0.27	0.63	0.27	0.43
VSXC	0.20	0.39	0.25	0.29
TPSSTPSS	0.21	0.46	0.25	0.33
O3LYP	0.51	0.66	0.27	0.52
B3LYP	0.51	0.66	0.27	0.52
B3P86	0.58	0.75	0.31	0.59
B3VP86	0.44	0.60	0.26	0.47
PBE1PBE	0.51	0.66	0.26	0.52
THCTHHYB	0.38	0.58	0.25	0.44
TPSSh	0.31	0.51	0.25	0.38
M05	0.67	0.86	0.30	0.67
BH&H	0.74	0.87	0.58	0.76
BH&HLYP	0.78	0.88	0.65	0.80
BMK	0.52	0.53	0.36	0.49
M05−2X	0.85	0.99	0.48	0.83
HSE1PBE	0.47	0.67	0.26	0.51
CAM-B3LYP	0.81	0.90	0.80	0.85
LC-BLYP	1.38	1.39	1.19	1.34
LC-ωPBE	0.95	1.06	0.86	0.98

In the alkene group, eleven transitions are considered for ethylene, two for isobutene, and seven for butadiene. The most intense states for these molecules are the first B_{1u} state

Table 4. X^L for the Azabenzenes

	pyridine	pyrazine	pyrimidine	pyridazine	S-tetrazine	all
RPA	2.44	1.65	1.96	2.75	1.02	1.78
CIS	2.79	2.14	2.26	2.94	1.44	2.22
CIS(D)	2.48	1.95	2.03	2.63	1.13	2.01
LSDA	1.02	0.94	1.45	0.65	0.67	0.99
BLYP	1.04	0.92	0.83	0.74	0.70	0.91
OLYP	0.83	0.70	0.52	0.77	0.67	0.69
BP86	1.02	0.98	1.14	0.92	0.71	0.99
BVP86	1.03	0.97	1.14	0.92	0.71	0.99
PBEPBE	1.03	0.95	0.56	0.77	0.67	0.90
HCTH	0.98	0.87	1.29	0.64	0.70	0.91
THCTH	1.12	0.97	1.23	0.73	0.74	1.00
VSXC	1.16	0.89	0.96	0.61	0.77	0.91
TPSSTPSS	1.05	0.74	1.11	0.93	0.73	0.81
O3LYP	1.41	1.16	1.42	1.14	0.68	1.20
B3LYP	1.39	1.17	1.30	1.21	0.74	1.20
B3P86	1.39	1.22	1.43	1.17	0.74	1.24
B3VP86	1.39	1.21	1.41	1.12	0.74	1.24
PBE1PBE	1.46	1.26	1.46	1.18	0.77	1.28
THCTHHYB	1.28	1.15	1.31	0.97	0.71	1.17
TPSSh	1.24	1.02	1.46	0.93	0.70	1.08
M05	1.67	1.39	1.69	1.19	0.66	1.42
BH&H	1.88	1.49	1.77	1.49	0.92	1.54
BH&HLYP	1.81	1.43	1.65	1.48	0.90	1.47
BMK	1.59	1.38	1.64	1.22	0.79	1.41
M05−2X	1.95	1.60	2.01	1.41	0.78	1.65
HSE1PBE	1.46	1.25	1.46	1.16	0.76	1.28
CAM-B3LYP	1.57	1.33	1.58	1.21	0.86	1.36
LC-BLYP	1.76	1.47	1.76	1.42	0.99	1.51
LC-ωPBE	1.66	1.42	1.73	1.31	0.96	1.46

for ethylene ($f^{\text{EOM-CCSD}}_{1B_{1u}} = 0.3520$), the A_1 state for isobutene ($f^{\text{EOM-CCSD}}_{A_1} = 0.1525$), and the first two B_u states for butadiene ($f^{\text{EOM-CCSD}}_{1B_u} = 0.6169$ and $f^{\text{EOM-CCSD}}_{2B_u} = 0.1636$). The oscillator strengths for the remaining states are on the order of 7×10^{-2} and lower. The WF methods overestimate the oscillator strength for this class of molecules, especially for isobutene. RPA, however, performs reasonably well for ethylene and butadiene (see Table 2). On the other hand, most of the density functionals underestimate the reference at least until a large percentage of HF exchange is introduced (see Table 2 and Figures 1 and 2). Hybrid functionals perform quite well for these molecules, although some provide a large overestimation for isobutene. This is the case for M05-2X, LC-BLYP, and LC-ωPBE, and to a lesser extent for B3P86, M05, BH&H, BH&HLYP, and CAM-B3LYP. The R^2 values are remarkably good with almost all of the methods for the individual molecules and as a group, as shown in Figures 3 and 4.

For the carbonyls, we consider eleven transitions for formaldehyde, six for acetaldehyde, and eight for acetone. The largest oscillator strengths are on the order of 7×10^{-2}. The WF methods largely overestimate the reference. CIS(D) reduces the value of X^L compared to CIS, but it is still large for acetone, see Table 3. In contrast, the DFT methods underestimate the oscillator strengths. The only exceptions are LC-BLYP for all of the molecules and LC-ωPBE for acetaldehyde, as shown in Table 3 and Figure 1. This underestimation is particularly severe for acetone in many

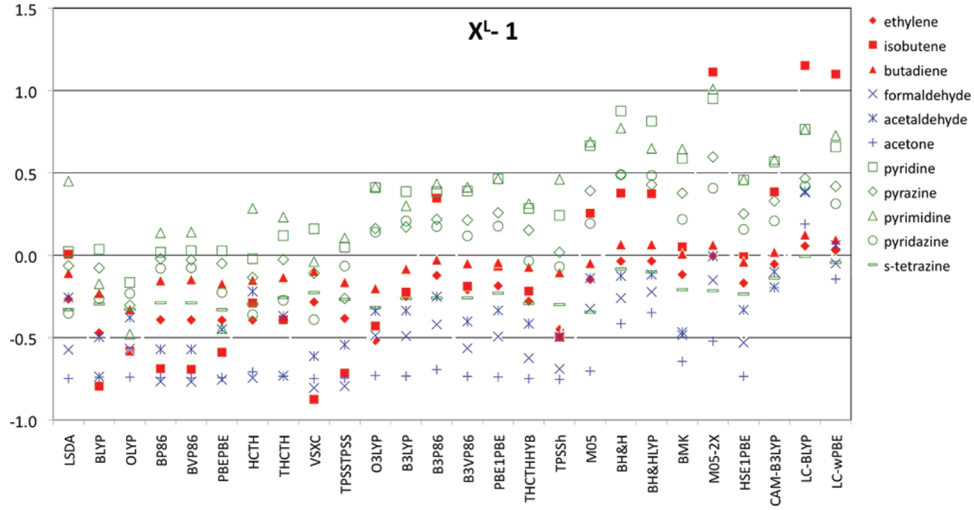


Figure 1. $X^L - 1$ for the single molecules.

cases. Very good performance is provided by LC-ωPBE and CAM-B3LYP, followed by BH&H and BH&HLYP. The R^2 values are much more scattered for this group, as shown in Figure 3. A general good behavior is reported for acetaldehyde, whereas poor results are obtained for formaldehyde and especially acetone with most of the pure and hybrid functionals with a small percentage of HF exchange (see Figure 3). CAM-B3LYP and LC-ωPBE provide a very good performance also for R^2, followed by BH&H and BH&HLYP. The X^L and R^2 values for the entire set (see Figures 2 and 4)

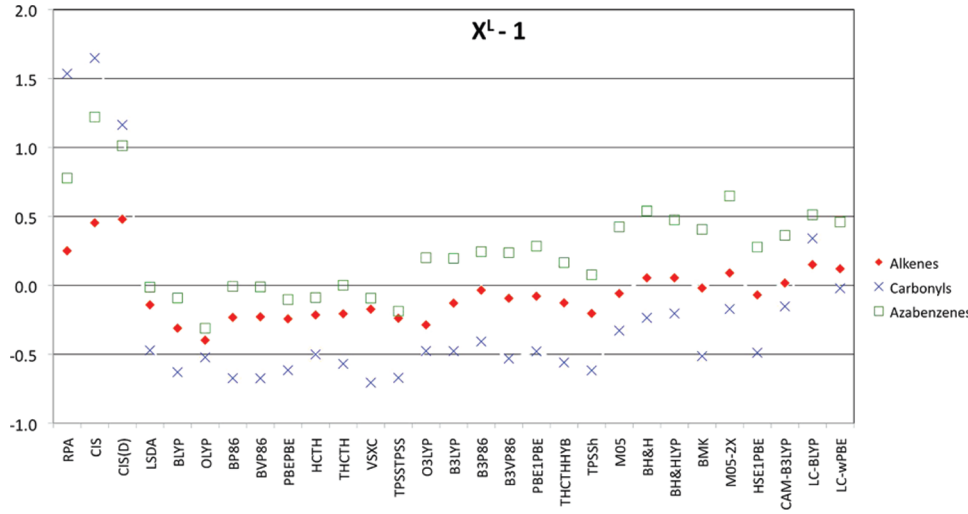


Figure 2. $X^L - 1$ for the molecular groups.

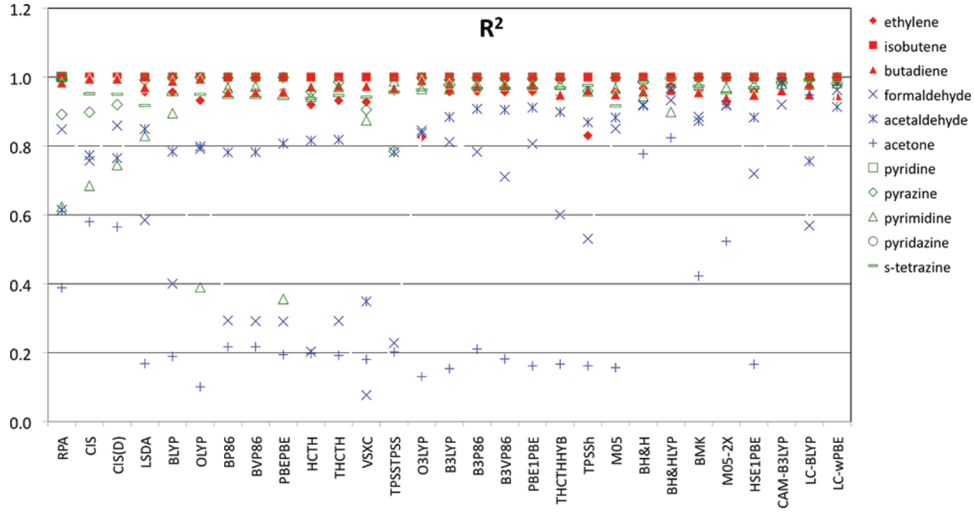


Figure 3. R^2 for the single molecules.

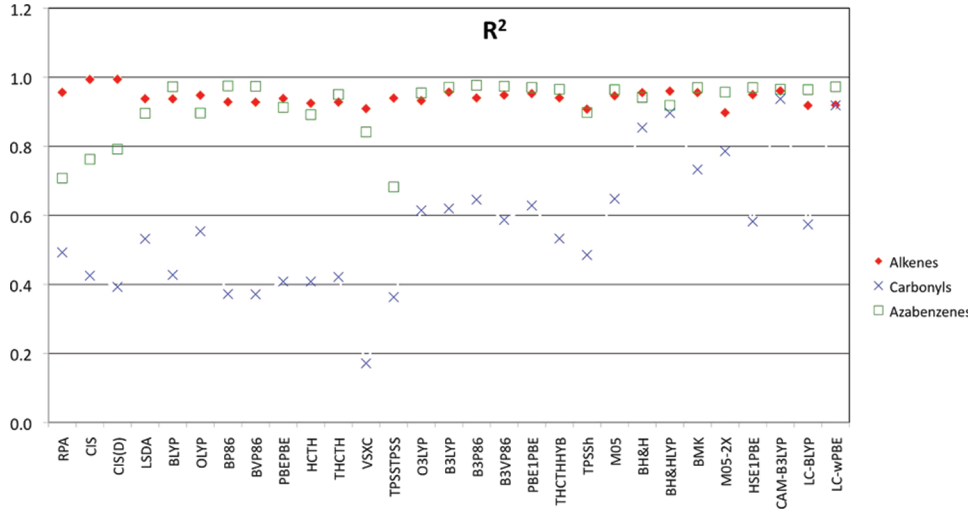


Figure 4. R^2 for the molecular groups.

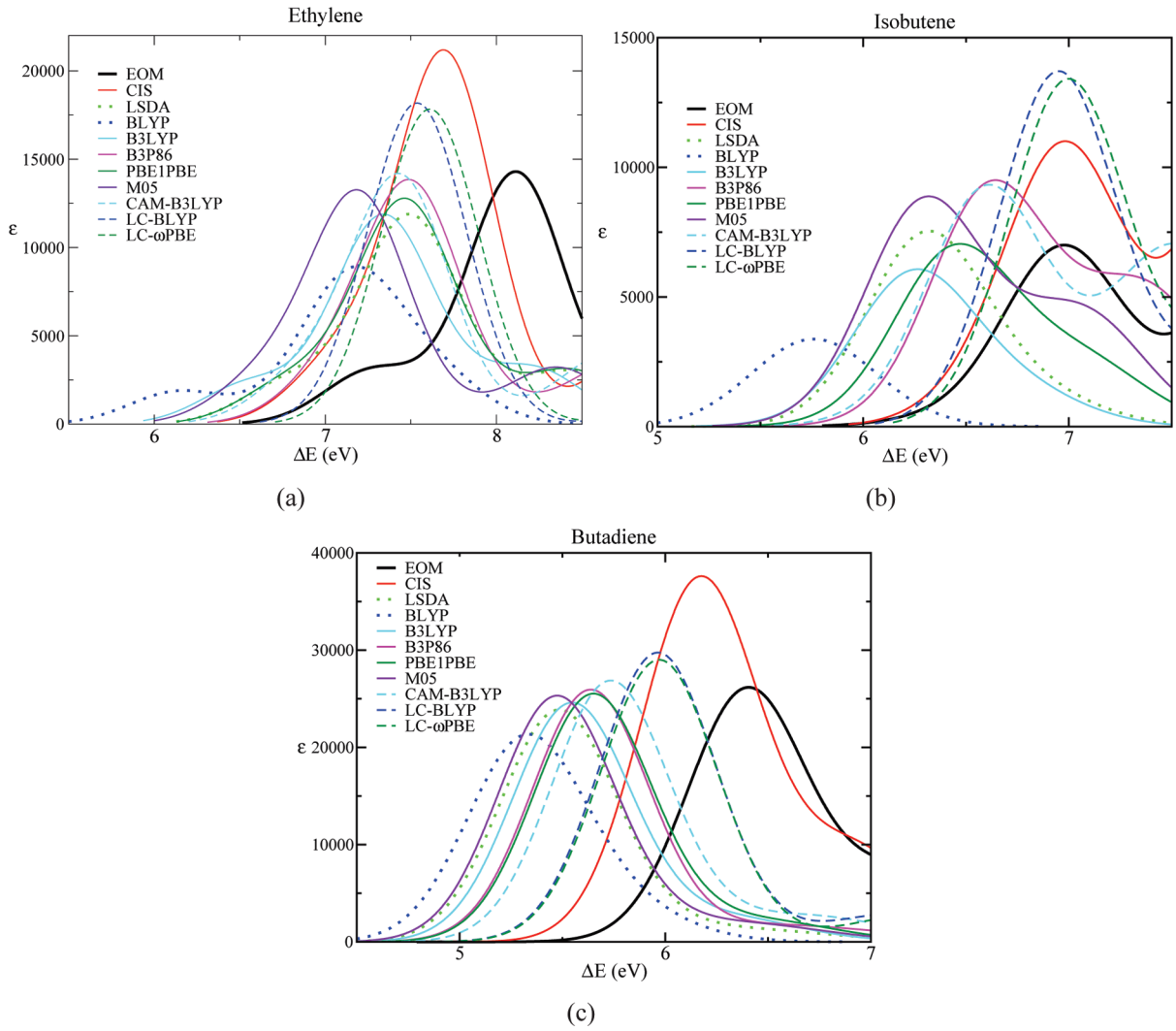


Figure 5. Alkenes spectra.

show a better performance moving from pure to hybrid functionals, especially those with large amounts of HF exchange. BH&H, BH&HLYP, CAM-B3LYP, and LC-ωPBE provide the overall best performance for this group.

In the azabenzene set, four transitions are considered for pyridine, five for pyrazine, six for pyrimidine, five for pyridazine, and four for s-tetrazine. The largest oscillator strengths are on the order of 8×10^{-2}. The X^L values are often larger than 2 for the WF methods, although RPA and CIS(D) provide rather good results for s-tetrazine (see Table 4). For the functionals, the magnitude of X^L seems to increase moving from left to right in Figure 1, i.e., going from pure functionals to hybrids with an increasing amount of HF exchange to functionals with short- and long-range separation. In this case, many pure functionals perform rather well, as do most of the hybrids with a small amount of HF exchange. CAM-B3LYP is the best among the long-range separated functionals. The R^2 values are reasonably good for almost all of the functionals with the exception of pyrimidine with OLYP and PBEPBE. The WF methods also show some difficulty with this molecule. The collective X^L results in Figure 2 show good performance of the pure functionals; this degrades for hybrid functionals, especially for those with a larger amount of HF exchange and short−long-range separation. On the other hand, the collective R^2 values in Figure 4 are very close to unity for most of the functionals.

4. Simulated Spectra

The spectrum of ethylene, Figure 5a, shows that all of the methods are shifted to lower energy compared to EOM-CCSD. CIS, LC-BLYP, and LC-ωPBE overestimate the intensity of the most intense band while the others are very close to the reference (only BLYP significantly underestimates it). Most of the DFT bands are centered around the same energy, while M05 and BLYP are shifted to lower energy. The small band centered at 7.3 eV for EOM-CCSD is not well reproduced by many methods, i.e., CIS, B3P86, M05, CAM-B3LYP, LC-BLYP, and LC-ωPBE. For these methods, the two bands are so close in energy that the large band covers the small one. For isobutene, Figure 5b, the CIS band is centered at the same energy of EOM-CCSD, but it is much more intense. The intensity of the band for LSDA, B3LYP, and PBE1PBE is similar to that for EOM-CCSD but shifted to lower energy. B3P86, M05, and CAM-B3LYP overestimate the intensity of the band, which is also shifted to lower energy. BLYP provides the worst performance with

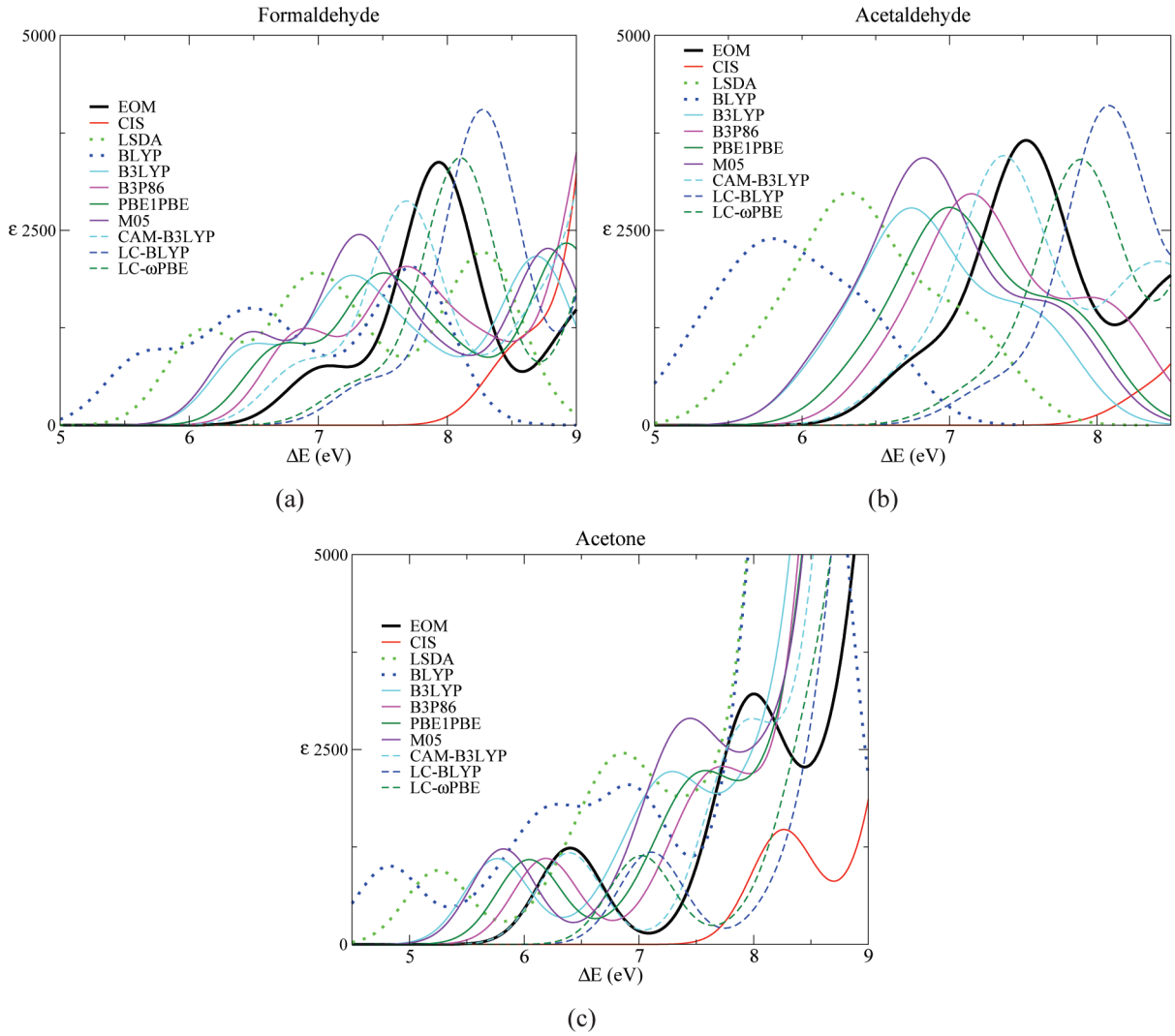


Figure 6. Carbonyls spectra.

a very small band at very low energy. LC-BLYP and LC-ωPBE have the band at the right energy, but the overestimation of its intensity is even greater than that for CIS. The butadiene spectrum, Figure 5c, is quite well reproduced by almost all of the functionals as far as the intensity of the band is concerned (the worst performance is again provided by BLYP), but the position of the bands is shifted to lower energy. The long-range corrected functionals overestimate the intensity of the band also in this case. CIS has the closest band to EOM-CCSD in terms of energy, but it is also the method that most overestimates the intensity.

The EOM-CCSD spectrum of formaldehyde has two bands at 7 and 8 eV (see Figure 6a). LSDA, BLYP, B3LYP, PBE1PBE, and M05 provide poor performance in this case, with bands of small intensity and shifted to lower energy. Poor performance is shown by CIS with very intense bands that are high in energy. CAM-B3LYP, LC-BLYP, and LC-ωPBE, on the other hand, satisfactorily reproduce the EOM-CCSD spectrum. CAM-B3LYP slightly underestimates the intensity of the band at 8 eV, while LC-BLYP and LC-ωPBE slightly overestimate it. B3P86 underestimates the intensity of the 8 eV band (similar to B3LYP and PBE1PBE), but its position is at the same energy as CAM-B3LYP. The spectrum of acetaldehyde, Figure 6b, is extremely overestimated by CIS both in the position and in the intensity of the bands (which are out of the energy range of the figure). The DFT methods do a better job with respect to the intensity, although only B3P86, CAM-B3LYP, and LC-ωPBE are close in energy to EOM-CCSD. The best agreement is provided by CAM-B3LYP with a spectrum slightly shifted to lower energy. For acetone, Figure 6c, CIS shows the usual large overestimation of the intensity and shifts to much higher energy. The opposite behavior is shown by LSDA and BLYP. The LC-BLYP and LC-ωPBE spectra are shifted to higher energy; these methods also miss the EOM-CCSD band at 8 eV. B3LYP, B3P86, and PBE1PBE similarly underestimate the intensity and position of the bands. Among these functionals, B3P86 is the closest to EOM-CCSD in terms of the energy. M05 provides a good performance for the intensity, but the bands are shifted to lower energy. The functional that best approximates the EOM-CCSD spectrum for this molecule in this energy range is CAM-B3LYP.

The azabenzenes have a small band at 5−5.5 eV and an intense band at 7.5−8 eV. For pyridine, Figure 7a, CIS overestimates the intensity of both, and their position is shifted to higher energy. All of the functionals simulate the EOM-CCSD spectrum more closely, although the intense

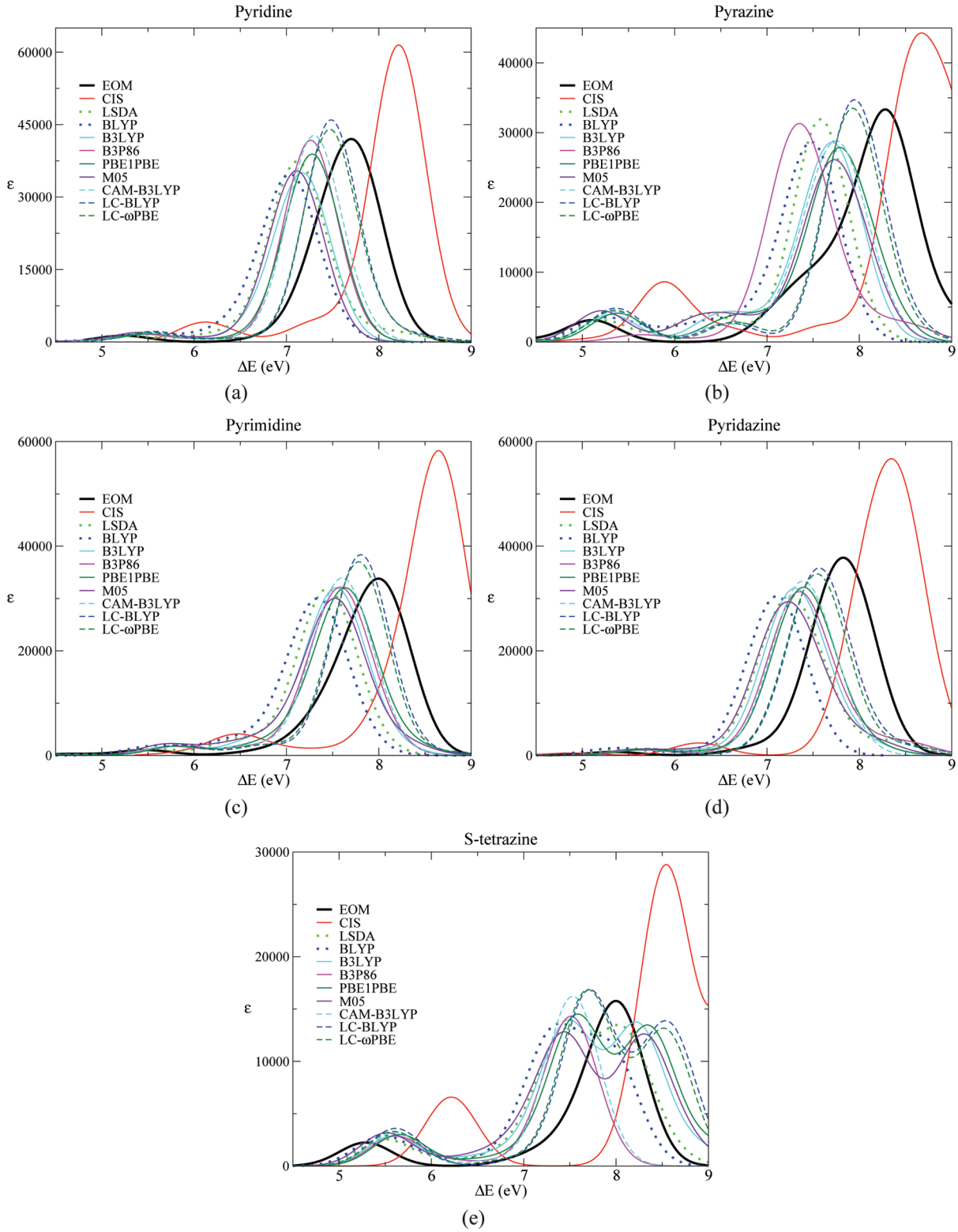


Figure 7. Azabenzenes spectra.

band is shifted to lower energy. For pyrazine, Figure 7b, the small band around 5 eV is shifted to higher energy with all of the methods. On the other hand, the more intense band is shifted to lower energy with all of the functionals (the CIS band is at higher energy). The best agreement with EOM-CCSD is obtained with LC-BLYP and LC-ωPBE, both in the intensity and in the position of the band. The other functionals underestimate the intensity of this band, although the intensities of the B3P86 and LSDA bands are close to that of EOM-CCSD. For pyrimidine, Figure 7c, similar considerations apply. However, in this case, the difference in the intensity of the large band is very small for most functionals. Only LC-BLYP and LC-ωPBE slightly overestimate the intensity of the EOM-CCSD band. CIS overestimates both bands in position and intensity. Pyridazine, Figure 7d, has a spectrum very similar to that of pyrimidine.

All of the methods behave comparably to the previous molecule, though the intensity of the large EOM-CCSD band is slightly underestimated by all of the functionals. Finally, s-tetrazine has a spectrum less intense in the same region (Figure 7e). LC-BLYP and LC-ωPBE reproduce the EOM-CCSD spectrum rather well, though the position of the small band is at higher energy while that of the larger band is at lower energy. The intensity of the large CAM-B3LYP band is comparable to that of EOM-CCSD, but it is centered at a slightly lower energy than those of LC-BLYP and LC-ωPBE. The third band shown by many functionals is outside the energy range considered for EOM-CCSD. Note that, with the exception of B3P86, CAM-B3LYP, LC-BLYP, and LC-ωPBE, the DFT methods require a large number of states to completely characterize the intense band for this set of molecules: around 30−40 states for the hybrid functionals and around 60−80 states for LSDA and BLYP.

5. Discussion and Conclusion

The results in sections 3 and 4 show that more approximate WF methods such as RPA, CIS, and CIS(D) consistently overestimate the oscillator strength compared to EOM-CCSD, as they do for the excitation energy. Additionally, the R^2 values show that the oscillator strengths are quite scattered compared to EOM-CCSD for the carbonyls and part of the azabenzenes. However, the CIS plots in section 4 qualitatively reproduce the EOM-CCSD spectra.

On the contrary, pure functionals mostly underestimate the oscillator strength (and the excitation energy[12]), in many cases dramatically, as is evident in Figure 1. Some good results are obtained for azabenzenes. The R^2 values in Figures 3 and 4 show good alignment for alkenes and azabenzenes but not for the carbonyls. The performance of LSDA and BLYP on the simulated spectra in section 4, where many states had to be included to completely characterize the most intense bands of the azabenzenes, suggests that several low lying Rydberg states appear. The underestimation of excitations to Rydberg states is a well-known problem of pure functionals,[41,66,83,84] which makes these functionals computationally expensive since many states need to be sought in order to include the ones of interest.

Hybrid functionals with a small percentage of HF exchange and no short−long-range separation tend to underestimate the oscillator strength for alkenes and carbonyls and overestimate it for azabenzenes. For these functionals, the R^2 values are close to unity for alkenes and azabenzenes but not for the carbonyl compounds, in particular, acetone. A larger amount of HF exchange shifts X^L (both individual and collective) to higher values, such that $X^L \simeq 1$ for the alkenes, $X^L < 1$ for the carbonyls, and $X^L > 1$ for the azabenzenes (see Figures 1 and 2). The R^2 parameters are closer to unity than for the pure functionals, but they are still rather small for the carbonyls, especially acetone. For the latter, a larger amount of HF exchange provides better results. The carbonyl set represents a difficult test for these functionals (and even more for pure functionals) since all of the relevant excitations considered in section 3 are to Rydberg states, whereas for the azabenzenes, all of the states are valence, and for the alkenes the very bright valence states are predominant. In the simulated spectra in section 4, we noticed a mixing of valence and Rydberg states, which leads to a redistribution of the intensity among several transitions as previously reported by Tozer et al.[10] However, this is hidden in the simulated spectra due to summing of the Gaussian line shapes. The simulated spectra in section 4 usually show less intense bands than EOM-CCSD with all of the functionals that do not separate short- and long-range effects. B3P86 stands out among these functionals as an excellent choice.

Three of the short−long-range separated functionals (CAM-B3LYP, LC-BLYP, and LC-ωPBE) perform on average better than the other functionals in reproducing the EOM-CCSD results, especially for difficult cases like acetone. This is shown in both the statistical analysis and the simulated spectra. In particular, CAM-B3LYP is often very close to the reference results, although it does not have the correct asymptotic behavior in the long range. LC-ωPBE also behaves well despite some difficulty with some azabenzenes and the significant overestimation of the isobutene oscillator strength for the bright A_1 state. CAM-B3LYP and LC-ωPBE also show the R^2 value closest to 1. In the simulated spectra, the two functionals with the correct asymptotic behavior, LC-BLYP and LC-ωPBE, often show more intense bands than EOM-CCSD.

For the data set we analyzed, i.e., Rydberg and valence states for small organic molecules, the magnitude of X^L increases moving from pure functionals to hybrids with more and more HF exchange. On the other hand, the WF methods in this work largely overestimate EOM-CCSD. More importantly, the results in Figures 1 and 2 seem to indicate that it is not possible to define a single scaling factor for the oscillator strength computed with the approximate functionals. In fact, $0.5 < X^L < 1.5$ depending on the set of molecules. On average, the best performance for the test cases considered here is obtained with CAM-B3LYP in both the statistical analysis and the simulation of the spectra, followed by LC-ωPBE. LC-BLYP also shows quite good agreement with EOM-CCSD. Among the functionals with no short−long-range separation, B3P86 provides satisfactory results with spectra often very close to those of CAM-B3LYP. As outlined in ref 12, the range of functionals and test cases considered in this work is certainly not complete. Nevertheless, these results provide useful insight on the ability of many current functionals to produce oscillator strengths of EOM-CCSD quality and show how much work is still necessary in the development of density functional theory.

Acknowledgment. M.C. thanks Prof. H. Bernhard Schlegel for encouraging this research and for stimulating discussions and Dr. Richard L. Lord for carefully reading the manuscript.

Supporting Information Available: Values of the oscillator strength and R^2 for all of the methods and the states in section 3. Extra diffuse functions added to the standard 6-311++G(d,p) basis to form the 6-311(3+,3+)G(d,p) basis set used in this work and in ref 12. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Casida, M. E. *Recent Advances in Density Functional Methods*; World Scientific: Singapore, 1995; Vol. 1.

(2) Casida, M. E. *Recent Developments and Applications of Modern Density Functional Theory, Theoretical and Computational Chemistry*; Elsevier: Amsterdam, 1996; Vol. 4.

(3) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.

(4) Wiberg, K. B.; de Oliveira, A. E.; Trucks, G. *J. Phys. Chem. A* **2002**, *106*, 4192–4199.

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

(6) Zhang, G.; Musgrave, C. B. *J. Phys. Chem. A* **2007**, *111*, 1554–1561.

(7) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.

(8) Silva-Junior, M. R.; Sauer, S. P. A.; Schreiber, M.; Thiel, W. *Mol. Phys.* **2010**, *108*, 453–465.

(9) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2010**, *133*, 174318.

(10) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andrés, L. *Mol. Phys.* **1999**, *97*, 859–868.

(11) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435, and references therein.

(12) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.

(13) Jacquemin, D. V.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Theor. Chem. Acc.* **2010**, *1*, 10 .

(14) Plötner, J.; Tozer, D. J.; Dreuw, A. *J. Chem. Theory Comput.* **2010**, *6*, 2315–2324.

(15) Jacquemin, D. V.; Perpete, E. A.; Ciofini, I.; Adamo, C.; Valero, R.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2071–2085.

(16) Jacquemin, D. V.; Perpète, E. A.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2010**, *6*, 1532–1537.

(17) Wiberg, K. B.; Hadad, C. M.; Ellison, G. B.; Foresman, J. B. *J. Phys. Chem.* **1993**, *97*, 13586–13597.

(18) Hadad, C. M.; Foresman, J. B.; Wiberg, K. B. *J. Phys. Chem.* **1993**, *97*, 4293–4312.

(19) Bolovinos, A.; Tsekeris, P.; Philis, J.; Pantos, E.; Andritsopoulos, G. *J. Mol. Spectrosc.* **1984**, *103*, 240–256.

(20) Walker, I. C.; Palmer, M. H.; Hopkirk, A. *Chem. Phys.* **1990**, *141*, 365–378.

(21) Goodman, L. *J. Mol. Spectrosc.* **1961**, *6*, 109–137.

(22) Innes, K. K.; Ross, I. G.; Moomaw, W. R. *J. Mol. Spectrosc.* **1988**, *132*, 492–544.

(23) Palmer, M. H.; Walker, I. C. *Chem. Phys.* **1991**, *157*, 187–200.

(24) Palmer, M. H.; McNab, H.; Reed, D.; Pollacchi, A.; Walker, I. C.; Guest, M. F.; Siggel, M. R. F. *Chem. Phys.* **1997**, *214*, 191–211.

(25) Spencer, G. H.; Cross, P. C.; Wiberg, K. B. *J. Chem. Phys.* **1961**, *35*, 1925–1938.

(26) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **2000**, *330*, 152–160.

(27) Nooijen, M. *J. Phys. Chem. A* **2000**, *104*, 4553–4561.

(28) Devarajan, A.; Gaenko, A. V.; Khait, Y. G.; Hoffmann, M. R. *J. Phys. Chem. A* **2008**, *112*, 2677–2682.

(29) For ethylene, we inverted the order of the two $1B_{1u}$ states in our previous work[12] for the following functionals: BLYP, OLYP, BP86, BVP86, PBEPBE, HCTH, THCTH, VSXC, TPSSTPSS, and TPSSh. Since the excitation energies for these two states are close to each other with these functionals, their final performance for this molecule is unchanged.

(30) Foresman, J. B.; Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *J. Phys. Chem.* **1992**, *96*, 135–149.

(31) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. *Chem. Phys. Lett.* **1994**, *219*, 21–29.

(32) McWeeny, R. *Methods of Molecular Quantum Mechanics*, 2nd ed.; Academic Press: London, 1992.

(33) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.

(34) Sekino, H.; Bartlett, R. J. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1984**, *18*, 255–265.

(35) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029–7039.

(36) Miura, M.; Aoki, Y.; Champagne, B. *J. Chem. Phys.* **2007**, *127*, 084103.

(37) Jacquemin, D.; Perpete, E. A. *THEOCHEM* **2007**, *804*, 31–34.

(38) Hirata, S.; Lee, T. J.; Head-Gordon, M. *J. Chem. Phys.* **1999**, *111*, 8904–8912.

(39) Tokura, S.; Tsuneda, T.; Hirao, K. *J. Theory Comput. Chem.* **2006**, *5*, 925–944.

(40) Matsuzawa, N. N.; Ishitani, A.; Dixon, D. A.; Uda, T. *J. Phys. Chem. A* **2001**, *105*, 4953–4962.

(41) Casida, M. E.; Salahub, D. R. *J. Chem. Phys.* **2000**, *113*, 8918–8935.

(42) Guillaumont, D.; Nakamura, S. *Dyes Pigment.* **2000**, *46*, 85–92.

(43) Peach, M. J. G.; Le Sueur, C. R.; Ruud, K.; Guillaumeb, M.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4465–4470.

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

(45) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

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

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

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

(49) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.

(50) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

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

(52) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400–410.

(53) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(54) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(55) Boese, A.; Doltsinis, N.; Handy, N.; Sprik, M. *J. Chem. Phys.* **2000**, *112*, 1670–1678.

(56) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497–5503.

(57) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559–9569.

(58) Hoe, W. M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319–328.

(59) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

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

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

(62) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405–3416.

(63) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207–8215.

(64) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(65) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(66) Tawada, T.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.

(67) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(68) Vydrov, O. A.; Heyd, J.; Krukau, V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.

(69) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(70) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

(71) Stephens, P. J.; Devlin, F. J.; Ashvar, C. S.; Chabalowski, C. F.; Frisch, M. J. *Faraday Discuss.* **1994**, *99*, 103–119.

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

(73) Staroverov, V. N.; Scuseria, G. E.; Tao, J. M.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

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

(75) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Scalmani, G.; Mennucci, B.; Barone, V.; Petersson, G. A.; Caricato, M.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Kudin, K. N.; Heyd, J. J.; Brothers, E.; Staroverov, V. N.; Zheng, G.; Kobayashi, R.; Normand, J.; Sonnenberg, J. L.; Ogliaro, F.; Bearpark, M.; Parandekar, P. V.; Ferguson, G. A.; Mayhall, N. J.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian Development Version*, Revision G.01+; Gaussian, Inc.: Wallingford, CT, 2008.

(76) Del Bene, J. E.; Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1997**, *106*, 6051–6060.

(77) Stanton, J. F.; Gauss, J.; Ishikawa, N.; Head-Gordon, M. *J. Chem. Phys.* **1995**, *103*, 4160–4174.

(78) Koch, H.; Jorgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333–3344.

(79) Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 9257–9269.

(80) Koch, H.; Kobayashi, R.; Demeras, A. S.; Jorgensen, P. *J. Chem. Phys.* **1994**, *100*, 4393–4400.

(81) Caricato, M.; Trucks, G. W.; Frisch, M. J. *J. Chem. Phys.* **2009**, *131*, 174104.

(82) Harada, N.; Chen, S. L.; Nakanishi, K. *J. Am. Chem. Soc.* **1975**, *97*, 5345–5352.

(83) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.

(84) Casida, M. E.; Casida, K. C.; Salahub, D. R. *Int. J. Quantum Chem.* **1998**, *70*, 933–941.

CT100662N

JCTC Journal of Chemical Theory and Computation

Anharmonicity and the Eigen-Zundel Dilemma in the IR Spectrum of the Protonated 21 Water Cluster

Miquel Torrent-Sucarrat* and Josep M. Anglada*

Departament de Química Biològica i Modelització Molecular, Institut de Química Avançada de Catalunya (IQAC−CSIC), c/Jordi Girona 18, E-08034 Barcelona, Spain

Received December 2, 2010

Abstract: The infrared anharmonic spectra for the $H^+(H_2O)_3$, $H^+(H_2O)_4$, and $H^+(H_2O)_{21}$ water clusters have been reported using vibrational second-order perturbation theory at the B3LYP level with 6-31+G(d) and 6-311++G(3df,3pd) basis sets. The anharmonicity results crucial for the evaluation of the protonated water clusters and the anharmonic corrections can be larger than 500 cm^{-1}, resulting in a shift of the H_3O^+ asymmetric stretchings near the region of 2000 cm^{-1}.

1. Introduction

Understanding the hydrated proton is of paramount importance for the knowledge of fundamental processes in chemistry and biology, and the investigation of protonated water clusters has been proven to be essential for understanding the nature of protons in solution.[1] From a conceptual point of view, the proton in the water clusters can be mainly located in two places. In the Eigen form (H_3O^+),[2] the proton is strongly linked by a single bond to the oxygen atom of a water molecule, while in the Zundel form,[3] it lies midway between the oxygen atoms of two water molecules ($H_2O-H^+-H_2O$). The Grotthuss mechanism[4] has been used to explain the proton transfer mechanism.

The structure and vibrational spectra of the protonated water clusters, $H^+(H_2O)_n$, have been a challenge for chemists in the recent decades. In the small size regime ($n \leq 11$), experimental and theoretical works have allowed characterization of the Eigen and Zundel motifs.[5−9] The structural identification of protonated water clusters with medium and large sizes has been more complicated. In 2004, as shown in the seminal works by Miyazaki et al.[6] and Shin et al.,[7] the authors were capable of isolating the protonated water clusters $H^+(H_2O)_n$ with $n = 6-27$ in gas phase and of measuring their infrared (IR) spectra from 2000 to 4000 cm^{-1}. Both works confirm that protonated water clusters are chains, two-dimensional nets, and three-dimensional cage structures at small, intermediate, and large sizes, respectively. Moreover, special attention has been paid to the vibrational spectrum of the $H^+(H_2O)_{21}$ cluster.

Mass spectroscopy studies by Lin[10] and Searcy and Fenn[11] found that $H^+(H_2O)_{21}$ shows a large mass peak intensity with respect to its neighboring clusters. This fact was ascribed to an exceptional stability of this cluster, and from that moment, $H^+(H_2O)_{21}$ was known as having the "magic number" of protonated water clusters. It was proposed that its stability is caused by its structure, and a distorted pentagonal dodecahedron cage, with a neutral water molecule encaged in the cavity, and a proton over the surface was suggested as an Eigen form. This hypothesis has been confirmed by theoretical calculations,[12,13] predicting that dodecahedral with an Eigen motif located at the cluster surface is the most stable conformer. Nevertheless, as far as we know, experimental confirmation of this hypothesis has not been reported yet.

In the infrared predissociation spectra (IRPD) series of $H^+(H_2O)_n$, with $n = 6-27$, obtained by Miyazaki et al.[6] and Shin et al.,[7] it was found that the O−H stretching band of two-coordinated single-acceptor−single-donor (AD) water molecules disappears at the magic number cluster. This fact is significant, because it supports the idea of a highly symmetric structure for $H^+(H_2O)_{21}$, i.e., a pentagonal dodecahedron cage with an internal water. However, these experimental results do not answer the question of whether the correct model for $H^+(H_2O)_{21}$ is an Eigen or a Zundel motif. The characteristic intense O−H stretching vibration band, predicted near 2500 cm^{-1}, for the Eigen structure does not appear in the experimental IR spectrum measured in the 2000−4000 cm^{-1} range.[6,7] This result supports the possibil-

* Phone: +34 934006111. Fax: +34 932045903. E-mail: mtsqbm@iqac.csic.es, anglada@iqac.csic.es.

10.1021/ct100692x
Published on Web 12/22/2010

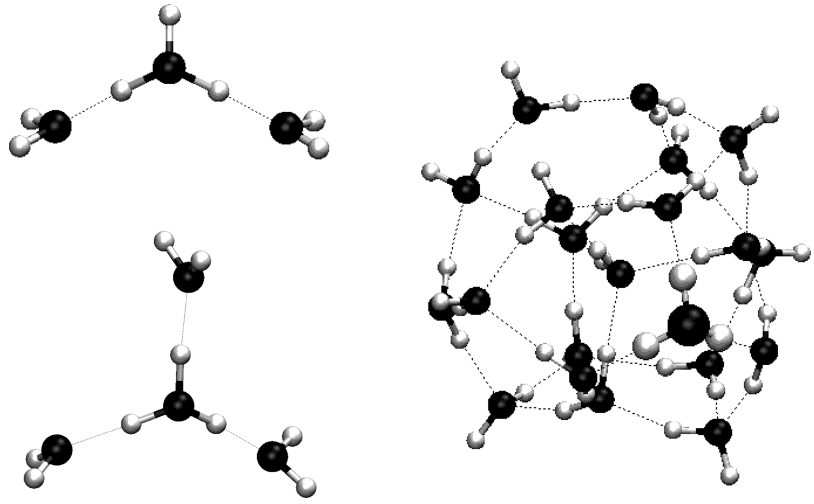


Figure 1. Optimized geometries for the protonated water clusters studied in this work. The atoms of the H_3O^+ cation in $H^+(H_2O)_{21}$ are enlarged.

ity of a Zundel structure. Nevertheless, the peculiar proton oscillation of the Zundel motif appears around 1100 cm^{-1}, which is below the experimental measured IR spectrum (in the 2000−4000 cm^{-1} range), and therefore, the Zundel motif cannot be confirmed.

The discrepancy between *ab initio* results and the experimental spectrum was addressed by the fact that the theoretical calculations do not include thermodynamic factors, like the effect of temperature[6,7] and the possible contributions from more than one structural isomer of a given cluster.[7,14] After these works, several authors[15] evaluated these thermodynamic factors for $H^+(H_2O)_{21}$, and it has been concluded that they have an important role in the vibrational spectrum. For instance, the nonappearance of the OH stretching vibrations for the hydronium in the experimental IRDP spectrum ranging from 2000 to 4000 cm^{-1} has been attributed to a reduction of their vibrational intensities due to thermodynamic effects.

In 2009, Duncan and co-workers[16] reported the IRPD spectra of $D^+(D_2O)_n$ (n = 18−24), including $H^+(H_2O)_{21}$, using a pulsed ring discharge source instead of the pulsed spark source previously employed.[7] The IRPD spectrum of $H^+(H_2O)_{21}$ is expanded to 1000 cm^{-1}, and in accordance with the previous results, the expected band around 2500 cm^{-1} of the O−H stretching vibrations of H_3O^+ was not detected. No other bands were found in the expanded lower frequency range of the spectrum. The pulsed ring discharge source makes larger ion signals, although the ions are not quite as cold as those produced with the laser spark. Then, in these new experiments, the role of the temperature becomes even more important. In addition, Douberly et al.[16] pointed out that dissociation energies or the dynamical rate of dissociation at lower energies might be the problem due to the lack of a signal of the IRPD spectrum below 3100 cm^{-1}, so that further IR spectroscopy of clusters at a well-defined cold temperature would be extremely valuable.

In this work, we present a different and complementary point of view of this problem. We calculate the IR anharmonic spectra for the most stable conformers of the $H^+(H_2O)_3$, $H^+(H_2O)_4$, and $H^+(H_2O)_{21}$ water clusters, all of them having an Eigen structure. In the incoming paragraphs, we will show that the O−H stretching vibrations of H_3O^+ have red-shifts larger than 500 cm^{-1}, and therefore anharmonicity can also play a crucial role in the characterization of the IR spectra of the $H^+(H_2O)_{21}$ water cluster.

2. Computational Methods

The B3LYP[17] exchange-correlation functional with 6-31+G(d) and 6-311++G(3df,3pd) basis sets[18] has been used to optimize the geometries of $H^+(H_2O)_3$, $H^+(H_2O)_4$, and $H^+(H_2O)_{21}$ (Figure 1), with the H_3O^+ sitting on the surface of the water cage, and to calculate their frequencies. All of the protonated water clusters of this work show an Eigen motif. The most stable conformer of $H^+(H_2O)_{21}$ with the hydronium ion sitting on the surface, obtained by Hodges and Wales,[12] has been used as the initial geometry of the optimization process.

In recent years, many different methodologies have been implemented to evaluate vibrational wave functions including anharmonicity. In the VSCF[19] procedure, each mode vibrates at the average potential generated by all other modes. The correlation between modes can be introduced through post-VSCF procedures such as perturbation theory (VMP2),[20] configuration interaction (VCI),[21] and coupled-cluster techniques (VCC).[22] The present work is focused on the vibrational second-order perturbation theory (VPT2)[23] treatment implemented by Barone.[24] The second-order perturbation theory correction is applied to a potential energy surface (PES) approximated by a Taylor series with normal coordinates, q_i, that includes the quadratic, all cubic, and semidiagonal quartic force constants.

$$V(q_1, q_2, ..., q_N) \cong \frac{1}{2}\sum_i w_i q_i^2 + \frac{1}{6}\sum_{ijk} f_{ijk} q_i q_j q_k + \frac{1}{24}\sum_{ijk} f_{ijkk} q_i q_j q_k q_k \quad (1)$$

The cubic and semidiagonal quartic force constants are computed using a finite difference approach, which linearly scales with the number of normal modes. For instance, in

Table 1. Calculated Harmonic and Anharmonic Vibrational Frequencies (in cm^{-1}) and Harmonic Intensities (in $km \cdot mol^{-1}$) at the B3LYP Level Using 6-311++G(3df,3pd) and 6-31+G(d) (in parentheses) Basis Sets for $H^+(H_2O)_3$

mode	*I* (har)	ν (har)	ν (anhar)	ν (exp)[a]
H_2O asym. stretching	374.7 (402.0)	3876.1 (3831.3)	3699.1 (3650.4)	3724
H_2O asym. stretching	2.2 (2.6)	3875.6 (3830.8)	3697.9 (3649.7)	3724
H_3O^+ free OH stretching	164.4 (145.7)	3801.4 (3737.4)	3616.3 (3546.8)	3580
H_2O sym. stretching	43.4 (74.3)	3787.6 (3728.8)	3620.5 (3559.4)	3639
H_2O sym. stretching	123.1 (121.8)	3787.1 (3729.0)	3621.4 (3564.1)	3639
H_3O^+ sym. stretching	1116.5 (1175.2)	2566.7 (2611.9)	2564.1 (2542.2)	2420
H_3O^+ asym. stretching	4137.9 (3990.3)	2382.0 (2440.8)	1802.1 (1848.6)	1880

[a] The experimental vibrational frequencies are obtained from ref 8, where $H^+(H_2O)_3$ has been tagged with Ar.

Table 2. Calculated Harmonic and Anharmonic Vibrational Frequencies (in cm^{-1}) and Harmonic Intensities (in $km \cdot mol^{-1}$) at the B3LYP Level Using 6-311++G(3df,3pd) and 6-31+G(d) (in parentheses) Basis Sets for $H^+(H_2O)_4$

Mode	*I* (har)	ν (har)	ν (anhar)	ν (exp)[a]
H_2O asym. stretching	402.9 (359.4)	3884.1 (3834.7)	3699.0 (3645.6)	3730
H_2O asym. stretching	37.1 (49.6)	3883.6 (3834.3)	3697.7 (3643.4)	3730
H_2O asym. stretching	32.3 (88.5)	3883.4 (3833.9)	3711.0 (3668.8)	3730
H_2O sym. stretching	1.4 (3.9)	3796.3 (3732.6)	3622.0 (3562.5)	3644
H_2O sym. stretching	74.1 (72.8)	3795.2 (3731.7)	3616.2 (3556.9)	3644
H_2O sym. stretching	73.6 (69.3)	3795.0 (3731.3)	3632.3 (3581.9)	3644
H_3O^+ asym. stretching	2994.0 (2853.3)	2856.7 (2869.0)	2622.5 (2609.1)	2665
H_3O^+ asym. stretching	2990.3 (2856.7)	2856.3 (2868.6)	2621.8 (2617.6)	2665

[a] The experimental vibrational frequencies are obtained from ref 8, where $H^+(H_2O)_4$ has been tagged with Ar.

the case of $H^+(H_2O)_{21}$, it implies 373 frequency calculations. The VPT2 method has been successfully applied to reproduce the vibrational properties for a large variety of systems.[25] All calculations have been carried out using Gaussian 03.[26]

3. Results and Discussion

The first step of the present work has been devoted to validating the VPT2 methodology used to evaluate the IR spectrum of $H^+(H_2O)_{21}$ with a potential energy surface (PES) obtained at the B3LYP/6-31+G(d) level.[27] To this end, we have computed the IR spectra of $H^+(H_2O)_3$ and $H^+(H_2O)_4$, for which accurate experimental IR spectra have been recently reported.[7,9] Tables 1 and 2 contain the experimental and calculated harmonic and anharmonic vibrational frequencies and harmonic intensities of selected vibrational frequencies for $H^+(H_2O)_3$ and $H^+(H_2O)_4$ systems, respectively, obtained at B3LYP level of theory using 6-31+G(d) and 6-311++G(3df,3pd) basis sets. The Supporting Information contains all of the harmonic and anharmonic vibrational modes for the protonated water clusters studied in this work.

The first important conclusion is that the augmentation of the basis set from 6-31+G(d) to 6-311++G(3df,3pd) is not crucial for the evaluation of the IR anharmonic spectra of $H^+(H_2O)_3$ and $H^+(H_2O)_4$ (for more details, see the Supporting Information). The second important conclusion is that the H_3O^+ stretchings present important anharmonic red-shifts, e.g., around 300 cm^{-1} for the $H^+(H_2O)_4$ system and up to 592.1 cm^{-1} for the asymmetric stretching of H_3O^+ for $H^+(H_2O)_3$.[28] Moreover, the introduction of anharmonicity through the VPT2 methodology produces a considerable improvement of the harmonic vibrational frequencies. The differences of the harmonic vibrational frequencies with respect to the experimental values range between 98.4 and 560.8 cm^{-1} for $H^+(H_2O)_3$ and between 88.6 and 204 cm^{-1} for $H^+(H_2O)_4$, while the differences between anharmonic and experimental vibrational frequencies range between 20.6 and 144.1 cm^{-1} for $H^+(H_2O)_3$ and between 22.0 and 84.4 cm^{-1} for $H^+(H_2O)_4$. These results illustrate the relevance of the anharmonicity in the evaluation of the vibrational spectra for protonated water clusters. In addition, we can conclude that the VPT2 methodology with the PES obtained at the B3LYP/6-31+G(d) level represents an effective and reliable choice for obtaining a semiquantitative reproduction of the experimental spectra of $H^+(H_2O)_{21}$.

Finally, it is interesting here to compare the IR spectra of both clusters, according to the results reported by Headrick et al.[8] In the case of the $H^+(H_2O)_3$ cluster, the Eigen signature appears at 1880 cm^{-1}. On the contrary, in the case of the $H^+(H_2O)_4$ cluster, the Eigen core is fully hydrated, and the corresponding band appears at 2665 cm^{-1}. In both cases, our anharmonic calculations predict quite well the experimental bands, and they are also in good agreement with the theoretical values reported in the same work. These results are very interesting because they clearly show that the Eigen signature can cover a wide range of frequencies. For instance, the OH stretch of the isolated H_3O^+ appears near 3500 cm^{-1},[29] and it is red-shifted up to 2665 cm^{-1} for the $H^+(H_2O)_4$ cluster and up to 1880 cm^{-1} for the $H^+(H_2O)_3$ and $H^+(H_2O)_5$ clusters.[8]

Regarding the $H^+(H_2O)_{21}$ cluster, Figure 2 shows the harmonic and anharmonic computed IR spectra, whereas the calculated values and relative intensities of the H_3O^+ stretchings are reported in Table 3.

At the first sight of Figure 2, one can see important dissimilarities between harmonic and anharmonic IR spectra. From 3000 to 4000 cm^{-1}, the symmetric and asymmetric stretchings of water molecules and the OH stretching bands of three-coordinated double-acceptor−single-donor (AAD) water molecules appear. These vibrational modes show different red-shift anharmonic corrections, ranging from 170 to 310 cm^{-1}.[28] These red-shift anharmonic corrections lead to a reorganization of the vibrational modes and an important

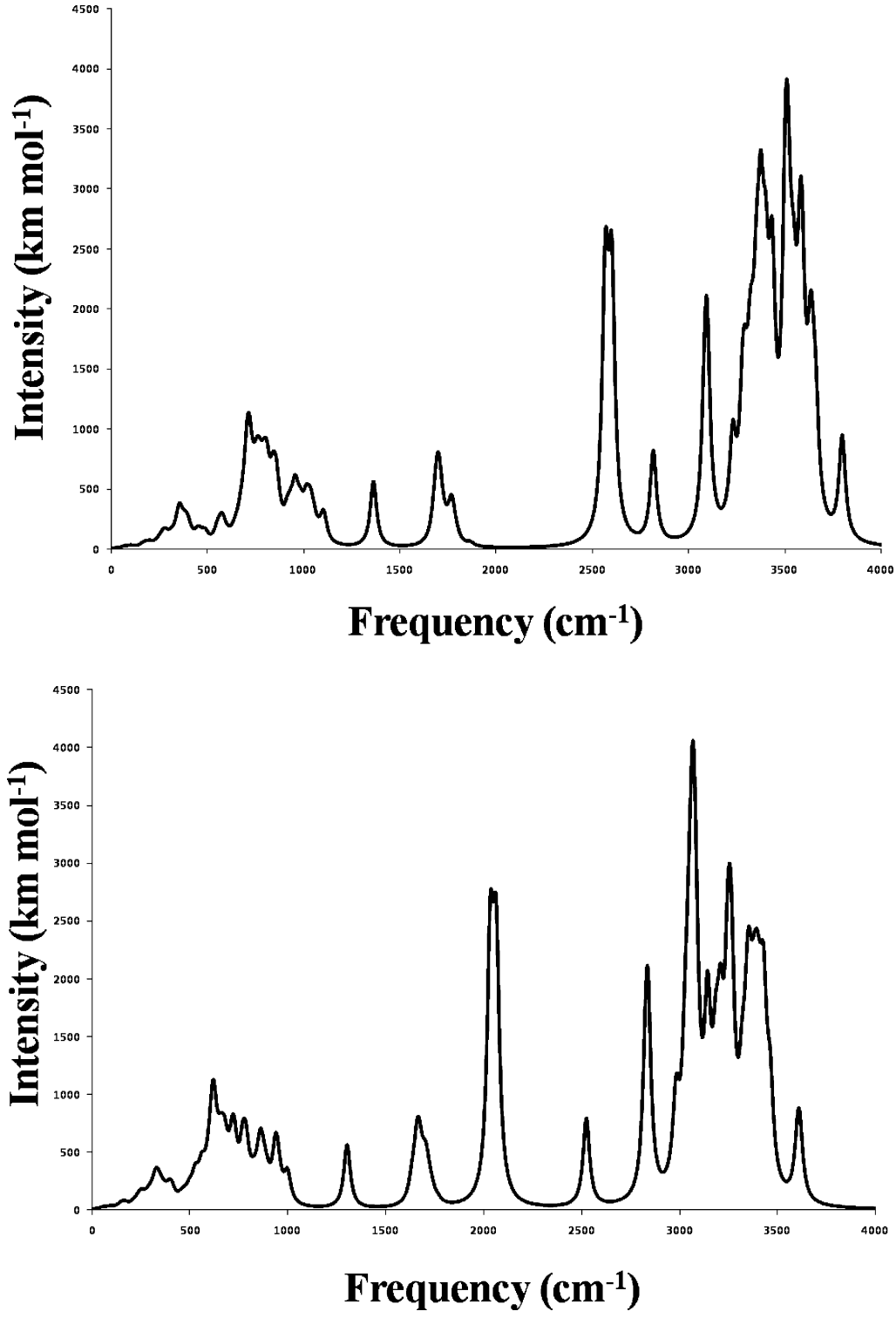


Figure 2. Computed harmonic (top) and anharmonic (bottom) spectra for the $H^+(H_2O)_{21}$ cluster with H_3O^+ located on the surface of the cage, obtained at the B3LYP/6-31+G(d) level of theory.

change of the IR spectrum shape, so that the shape of the anharmonic computed IR spectrum of the $H^+(H_2O)_{21}$ cluster is more similar to the experimental IRPD spectrum reported by Shin et al.[7] and Douberly et al.[16] than that predicted by the harmonic approach.

Below 3000 cm^{-1}, the "problematic" stretching bands of H_3O^+ are found. First, our calculations predict a less intense band (761 $km \cdot mol^{-1}$), which is located at 2523.3 cm^{-1}, that corresponds to the H_3O^+symmetric stretching. This band has an important anharmonic correction of 292.5 cm^{-1}. Second, our calculations predict at 2062.9 and 2033.2 cm^{-1} two intense bands (2030 and 2080 $km \cdot mol^{-1}$), corresponding to the fingerprint vibrational frequencies of the H_3O^+ asymmetric stretchings. It is important to remark here that, for these two bands, the anharmonic corrections are larger than 500 cm^{-1} in a similar way to that previously discussed for the $H^+(H_2O)_3$ cluster. Finally, below 1800 cm^{-1}, the harmonic and anharmonic vibrational IR spectra are quite

Table 3. Calculated Harmonic and Anharmonic Vibrational Frequencies (in cm^{-1}) and Harmonic Intensities (in $km \cdot mol^{-1}$) at the B3LYP/6-31+G(d) Level of H_3O^+ Stretching for the $H^+(H_2O)_{21}$ Cluster with H_3O^+ Located on the Surface of the Cage

mode	*I* (har)	ν (har)	ν (anhar)
H_3O^+ sym. stretching	761.4	2815.8	2523.3
H_3O^+ asym. stretching	2030.4	2599.8	2062.9
H_3O^+ asym. stretching	2079.9	2567.2	2033.2

analogous, and they contain vibrational bending frequencies of water molecules with anharmonic corrections smaller than 100 cm^{-1}.

Our predicted anharmonic spectrum shows large differences with respect to the experimental IRPD spectra reported in the literature[6,7,16] in which neither the H_3O^+ stretching frequencies nor the vibrational bending frequencies below 1800 cm^{-1} are observed. According to previous works,[7,14−16] these differences might be attributed to thermodynamic and dynamic effects, like the temperature and the contributions from more than one structural isomer of a given cluster. In addition, it is worth pointing out here that the calculations and experiments address different situations; namely, the calculations correspond to one-photon absorption spectra, and the experiment corresponds to multiphoton (predominantly two-photon) predissociation spectra. These factors can have an important role in the vibrational spectrum of $H^+(H_2O)_{21}$, and consequently, they can lead to a reduction of the vibrational band intensities. For instance, they could explain the nonappearances of the symmetric H_3O^+ stretching vibration predicted around 2500 cm^{-1} and the asymmetric H_3O^+ stretching vibrations predicted around 2000 cm^{-1} in the experimental IR spectra. Our computed IR spectrum does not consider the effect of the temperature, and it corresponds to the most stable conformation of the $H^+(H_2O)_{21}$ cluster. Thus, our results should mimic a spectrum measured at very low temperatures, where the thermodynamic and dynamic effects are reduced.

4. Conclusions

In the present work, we have studied the anharmonic effects of the $H^+(H_2O)_3$, $H^+(H_2O)_4$, and $H^+(H_2O)_{21}$ water clusters, all of them having an Eigen structure, by using the VPT2 methodology with a PES obtained at the B3LYP level of theory using 6-31+G(d) and 6-311++G(3df,3pd) basis sets. Our results lead to the following conclusions:

For the $H^+(H_2O)_3$ and $H^+(H_2O)_4$ clusters, experimental and other theoretical data are available, and they are in good agreement with the results obtained in this work. According to our calculations, the Eigen signature appears at 1802 cm^{-1} in the case of the $H^+(H_2O)_3$ cluster and at 2622 cm^{-1} in the case of the $H^+(H_2O)_4$ cluster (the experimental values are 1880 and 2665 cm^{-1}, respectively). In both cases, the anharmonic effects are shown to be very important, with red-shifts up to 580 cm^{-1} for $H^+(H_2O)_3$ and up to 234 cm^{-1} for $H^+(H_2O)_4$ with respect to the harmonic spectra.

Regarding to the $H^+(H_2O)_{21}$ cluster, our calculations show that the anharmonic effects produce important red-shifts in the computed bands. In the 3000 to 4000 cm^{-1} range, the computed anharmonic IR spectrum is more similar to the experimental spectrum than the harmonic one. Moreover, our calculations predict the symmetric and HO asymmetric stretchings of the Eigen signature to appear close to 2500 and 2000 cm^{-1} with red-shifts larger than 250 and 500 cm^{-1}, respectively, with respect to the computed harmonic value. The discrepancies between theoretical and experimental IR spectra have been attributed to thermodynamic and dynamic effects, and therefore, these results will be valid at very low temperatures. Regarding this point, and following a reviewer's comment, it is worth noting here that it will be very difficult to obtain a high enough concentration of ions to do a direct IR adsorption measurement, and at temperatures near 0 K, the IRPD technique would take more than two photons to observe dissociation on the time scale of the experiment.

Acknowledgment. We thank the reviewers for their comments and suggestions regarding this work. This research has been supported by the Spanish Dirección General de Investigación Científica y Técnica (DGYCIT, grant CTQ2008-06536/BQU), the Generalitat de Catalunya (Grant 2009-SGR01472), and the Research Executive Agency (Grant Agreement no. PERG05-GA-2009-249310). The calculations described in this work were carried out at the Centre de Supercomputació de Catalunya (CESCA) and at the CTI-CSIC. M.T-S. acknowledges the CSIC for the JAE-DOC contract.

Supporting Information Available: All of the harmonic and anharmonic vibrational modes at the B3LYP level of theory using 6-31+G(d) and 6-311++G(3df,3pd) basis sets for the protonated water clusters studied in this work. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Kunst, M.; Warman, J. M. *Nature* **1980**, *288*, 465. Heberle, J.; Riesle, J.; Thiedemann, G.; Oesterhelt, D.; Dencher, N. A. *Nature* **1994**, *370*, 379. Waldmann, R.; Champigny, G.; Bassilana, F.; Heurteaux, C.; Lazdunski, M. *Nature* **1997**, *386*, 173. Luecke, H.; Richter, H. T.; Lanyi, J. K. *Science* **1998**, *280*, 1934. Marx, D.; Tuckerman, M. E.; Hutter, J.; Parrinello, M. *Nature* **1999**, *397*, 601. Bakker, H. J.; Nienhuys, H. K. *Science* **2002**, *297*, 587. Ludwig, R. *Chemphyschem* **2004**, *5*, 1495. Sun, Z.; Siu, C. K.; Balaj, O. P.; Gruber, M.; Bondybey, V. E.; Beyer, M. K. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 4027.

(2) Eigen, M. *Angew. Chem., Int. Ed. Engl.* **1964**, *3*, 1.

(3) Zundel, G.; Metzger, H. *Z. Phys. Chem.* **1968**, *58*, 225.

(4) Agmon, N. *Chem. Phys. Lett.* **1995**, *244*, 456.

(5) Jiang, J. C.; Wang, Y. S.; Chang, H. C.; Lin, S. H.; Lee, Y. T.; Niedner-Schatteburg, G. *J. Am. Chem. Soc.* **2000**, *122*, 1398. Asmis, K. R.; Pivonka, N. L.; Santambrogio, G.; Brummer, M.; Kaposta, C.; Neumark, D. M.; Woste, L. *Science* **2003**, *299*, 1375. Vendrell, O.; Meyer, H. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4692. Nguyen, Q. C.; Ong, Y. S.; Kuo, J. L. *J. Chem. Theory Comput.* **2009**, *5*, 2629. Karthikeyan, S.; Kim, K. S. *J. Phys. Chem. A* **2009**, *113*, 9237. Vendrell, O.; Gatti, F.; Meyer, H. D. *Angew. Chem., Int. Ed. Engl.* **2009**, *48*, 352. Kaledin, M.; Wood, C. A. *J. Chem. Theory Comput.* **2010**, *6*, 2525.

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

(7) Shin, J. W.; Hammer, N. I.; Diken, E. G.; Johnson, M. A.; Walters, R. S.; Jaeger, T. D.; Duncan, M. A.; Christie, R. A.; Jordan, K. D. *Science* **2004**, *304*, 1137.

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

(9) Douberly, G. E.; Walters, R. S.; Cui, J.; Jordan, K. D.; Duncan, M. A. *J. Phys. Chem. A* **2010**, *114*, 4570.

(10) Lin, S. S. *Rev. Sci. Instrum.* **1973**, *44*, 516.

(11) Searcy, J. Q.; Fenn, J. B. *J. Chem. Phys.* **1974**, *61*, 5282.

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

(13) Mrazek, J.; Burda, J. V. *J. Chem. Phys.* **2006**, *125*, 194518. Kus, T.; Lotrich, V. F.; Perera, A.; Bartlett, R. J. *J. Chem. Phys.* **2009**, *131*, 104313.

(14) Zwier, T. S. *Science* **2004**, *304*, 1119.

(15) Iyengar, S. S.; Petersen, M. K.; Day, T. J. F.; Burnham, C. J.; Teige, V. E.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 084309. Wu, C. C.; Lin, C. K.; Chang, H. C.; Jiang, J. C.; Kuo, J. L.; Klein, M. L. *J. Chem. Phys.* **2005**, *122*, 074315. Singh, N. J.; Park, M.; Min, S. K.; Suh, S. B.; Kim, K. S. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 3795. Iyengar, S. S. *J. Chem. Phys.* **2007**, *126*, 216101.

(16) Douberly, G. E.; Ricks, A. M.; Duncan, M. A. *J. Phys. Chem. A* **2009**, *113*, 8449.

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

(18) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986; pp 113−225.

(19) Bowman, J. M. *J. Chem. Phys.* **1978**, *68*, 608. Gerber, R. B.; Ratner, M. A. *Chem. Phys. Lett.* **1979**, *68*, 195. Bowman, J. M. *Acc. Chem. Res.* **1986**, *19*, 202. Gerber, R. B.; Ratner, M. A. *Adv. Chem. Phys.* **1988**, *70*, 97.

(20) Jung, J.-Q.; Gerber, R. B. *J. Chem. Phys.* **1996**, *105*, 10332. Christiansen, O. *J. Chem. Phys.* **2003**, *119*, 5773. Bowman, J. M.; Carter, S.; Huang, X.-C. *Int. Rev. Phys. Chem.* **2003**, *22*, 533.

(21) Bowman, J. M.; Christoffel, K. M.; Tobin, F. *J. Phys. Chem.* **1979**, *83*, 905. Christoffel, K. M.; Bowman, J. M. *Chem. Phys. Lett.* **1982**, *85*, 220. Carter, S.; Bowman, J. M.; Handy, N. C. *Theor. Chim. Acta* **1998**, *100*, 191.

(22) Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2149. Christiansen, O. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2942.

(23) Darling, B. T.; Dennison, D. M. *Phys. Rev.* **1940**, *57*, 128. Wilson, J. E. B.; Decius, J. C.; Cross, P. C. *Molecular vibrations, The theory of infrared and Raman Vibrational Spectra*; Dover: New York, 1955; pp 169−202. Clabo, D. A.; Allen, W. D.; Remington, R. B.; Yamaguchi, Y.; Schaefer, H. F., III. *Chem. Phys.* **1988**, *123*, 187. Halonen, L. *J. Chem. Phys.* **1997**, *106*, 7931. Hanninen, V.; Horn, M.; Halonen, L. *J. Chem. Phys.* **1999**, *111*, 3018. Dreyer, J. *J. Chem. Phys.* **2007**, *127*, 054309.

(24) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059. Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108.

(25) Carbonniere, P.; Lucca, T.; Pouchan, C.; Rega, N.; Barone, V. *J. Comput. Chem.* **2005**, *26*, 384. Dunn, M. E.; Evans, T. M.; Kirschner, K. N.; Shields, G. C. *J. Phys. Chem. A* **2006**, *110*, 303. Yavuz, I.; Trindle, C. *J. Chem. Theory Comput.* **2008**, *4*, 533. Torrent-Sucarrat, M.; Anglada, J. M.; Luis, J. M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6377. Cappelli, C.; Monti, S.; Scalmani, G.; Barone, V. *J. Chem. Theory Comput.* **2010**, *6*, 1660. Figgen, D.; Koers, A.; Schwerdtfeger, P. *Angew. Chem., Int. Ed. Engl.* **2010**, *49*, 2941. Marta, R. A.; Wu, R. H.; Eldridge, K. R.; Martens, J. K.; McMahon, T. B. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3431.

(26) 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, 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 ed.; Gaussian, Inc.: Pittsburgh, PA, 2003.

(27) It is important to remark that the harmonic and anharmonic vibrational spectra for the $H^+(H_2O)_3$ and $H^+(H_2O)_4$ clusters were also evaluated at the MP2 level, and they compare quite well with the B3LYP and the experimental results.

(28) It is important to remark that some of the largest anharmonic corrections to vibrational frequencies may suffer from some errors. They have been evaluated using a trucanted PES and the perturbation theory. The coupling between anharmonic low and high frequency modes can generate very large cubic and quartic force constants, which could affect the convergence of the perturbation method. For instance, in the low frequency modes of $H^+(H_2O)_3$ and $H^+(H_2O)_4$, the VPT2 correction leads to negative vibrational frequencies. In order to fix these modes and to validate the VPT2 results for the highly anharmonic systems, it would be necessary to apply vibrational variational methods, i.e., vibrational configuration interaction (VCI) or vibrational coupled-cluster (VCC), and to use a full PES, although this point is outside the scope and possibilities of this work.

(29) Pribble, R. N.; Zwier, T. S. *Science* **1994**, *265*, 75.

CT100692X

JCTC Journal of Chemical Theory and Computation

A Classical Potential to Model the Adsorption of Biological Molecules on Oxidized Titanium Surfaces

Julian Schneider*,† and Lucio Colombi Ciacchi†,‡

Hybrid Materials Interfaces Group, Faculty of Production Engineering and Bremen Center for Computational Materials Science, University of Bremen, D-28359 Bremen, Germany and Fraunhofer Institute for Manufacturing Technology and Applied Materials Research IFAM, D-28359 Bremen, Germany

Received August 6, 2010

Abstract: The behavior of titanium implants in physiological environments is governed by the thin oxide layer that forms spontaneously on the metal surface and mediates the interactions with adsorbate molecules. In order to study the adsorption of biomolecules on titanium in a realistic fashion, we first build up a model of an oxidized Ti surface in contact with liquid water by means of extensive first-principles molecular dynamics simulations. Taking the obtained structure as reference, we then develop a classical potential to model the Ti/TiO_x/water interface. This is based on the mapping with Coulomb and Lennard-Jones potentials of the adsorption energy landscape of single water and ammonia molecules on the rutile TiO_2(110) surface. The interactions with arbitrary organic molecules are obtained via standard combination rules to established biomolecular force fields. The transferability of our potential to the case of organic molecules adsorbing on the oxidized Ti surface is checked by comparing the classical potential energy surfaces of representative systems to quantum mechanical results at the level of density functional theory. Moreover, we calculate the heat of immersion of the TiO_2 rutile surface and the detachment force of a single tyrosine residue from steered molecular dynamics simulations, finding good agreement with experimental reference data in both cases. As a first application, we study the adsorption behavior of the Arg-Gly-Asp (RGD) peptide on the oxidized titanium surface, focusing particularly on the calculation of the free energy of desorption.

1. Introduction

The outstanding mechanical and chemical properties of titanium have attracted for decades the attention of materials scientists, leading to the development of Ti-based alloys for a broad range of applications. Besides its wide use in the aerospace and marine industries, its corrosion resistance and biocompatibility make titanium a material of choice for medical and dental implants.[1,2] In this case, a thorough knowledge of the physical and chemical details of the interface between the implant and the physiological environment is desired for tailoring the surface properties and optimizing the adhesion of cells within the body tissues. Since these processes are governed by the adsorption of biological macromolecules, such as proteins, an atomic-scale understanding of the interaction between proteins and the metal surface is often sought, yet still lacking.[3]

Complementary to experiments, atomistic molecular dynamics (MD) simulations, based on either quantum mechanical or classical formalisms, may provide a powerful method to gain insight into the microscopic mechanisms involved in protein adhesion. However, realistic simulations of the interface between titanium and a physiological environment have to face the rich chemical complexity of the system, which prevents the use of simple structural models and generic interaction potentials. In contact with water and oxygen, the metallic Ti surface is covered by an oxide layer, whose composition, structure, and thickness strongly depend on the oxidation conditions. It is known that high temper-

* Corresponding author. E-mail: schneider@hmi.uni-bremen.de.
† University of Bremen.
‡ Fraunhofer IFAM.

10.1021/ct1004388
Published on Web 12/23/2010

atures promote the formation of a thick TiO_2 layer, whereas oxidation at room temperature results in thin layers ($\leq$1 nm) composed of a broad range of titanium oxidation states and with stoichiometries variable from Ti_2O to TiO_2.[4−9]

When considering molecular adsorption on Ti surfaces, it is crucial to take into account the precise structure and chemistry of the oxide layer. In particular, since the oxidized Ti surface does not reveal single crystal features, it might be a too strong approximation to model it with an ideal TiO_2 crystal facet, as done so far in many simulation studies.[10−15] Recently, in extensive quantum mechanical MD simulations of the oxidation reactions of the bare metal, we have obtained a realistic model for the oxidized Ti(0001) surface, including two monolayers (ML) of chemisorbed oxygen atoms.[16] In agreement with the available experimental knowledge, this model reveals a rather amorphous oxide structure and variable Ti oxidation states and thus appears to capture well the representative features of Ti surfaces exposed to an oxidizing environment.

Although the adsorption reactions of small molecules, such as oxygen, can be simulated accurately by means of quantum mechanics,[11,17,18] the adsorption of large molecules over correspondingly extended surface areas can nowadays be investigated only by means of classical simulations. These require suitable 'force field' potentials to model the surface dynamics and the interactions between surface and adsorbate molecules. Several approaches are currently in use to model the interactions at the interfaces between titania and water or solvated organic molecules.[12,19,12] However, the current models are based on perfect crystal surfaces, and their transferability to thin oxide layers including various Ti oxidation states is uncertain. Furthermore, the available potentials have been scarcely validated for systems including more than 1 ML of water molecules, and the interactions with bulk water are rarely tested against suitable experimental results, such as the heat of immersion. Also the applicability of combination rules to extend the potentials to arbitrary molecules is often assumed but not investigated carefully.

As reported in ref 16, we have recently developed a force field to model the dry oxidized titanium surface in classical MD simulations. In this paper we present an extension of this force field to model Ti/TiO_x/water interfaces and in particular solvated organic adsorbate molecules interacting with oxidized Ti. The simple analytical form of our potential, based on atomic point charges and Lennard-Jones (LJ) interactions, is compatible with commonly used water models and biomolecular force fields and makes feasible the simulation of large systems. The potential parameters are accurately tuned on the basis of density functional theory (DFT) calculations of the potential energy surfaces (PES) of various organic molecules on the dry surface and validated against experimental results for the heat of immersion of TiO_2 as well as the adhesion force of tyrosine on oxidized Ti. As a first application, we present simulations of the adsorption behavior of the amino acid sequence Arg-Gly-Asp (RGD), which is widely used to functionalize biomaterials surfaces with the aim of promoting a better surface adhesion of cells in biomedical implants.

The paper is structured as follows: After a summary of the computational methods (Section 2), in Section 3 we describe first-principles molecular dynamics (FPMD) simulations of the interfaces between bulk water and both TiO_2 and the oxidized Ti(0001) surface. These calculations are used as a reference model for the construction of our potential in Section 4. In particular, we focus on an analysis of the charges of surface atoms and on how to achieve consistency with generic biomolecular force fields. We then proceed to derive appropriate nonbonded interactions and optimal parameters from DFT calculations of the PES of water and ammonia on the TiO_2 surface. In Section 4.4 we compare the classical model to DFT results of small organic molecules adsorbed on the dry oxidized surface. Subsequently, classical simulations of wet systems are presented and discussed in comparison to experimental results in Section 5. Finally, the adsorption behavior of the RGD peptide on the oxidized titanium surface is investigated in Section 6.

2. Computational Details

2.1. FPMD Simulations. Our FPMD simulations are performed within the formalism of DFT, employing the PW91 exchange correlation GGA functional[22] and the projector-augmented wave (PAW) method[23] to represent the interactions between electrons and core ions, as implemented in the Lautrec code.[24] The PAW data set for Ti is generated with 12 explicit valence electrons, including 3, 2, 2 projectors for the s, p, and d angular momentum channels. The data sets for O, N, and C include six, five, and four valence electrons and two projectors in each of the s and p channels. The wave functions are expanded in plane waves up to a kinetic energy cutoff of 540 eV. Since all systems under investigation are nonmagnetic, we employ spin-paired calculations. When considering metallic systems the electronic states are occupied according to a Fermi-Dirac distribution using a smearing width of 0.1 eV. Both the minimization of the electronic states and the MD simulations are performed using the Car−Parrinello (CP) method,[25] making use of special algorithms for the treatment of metallic systems[26,27] where necessary. The surface cells are sampled using the (0.25, 0.25) point of the Brillouin zone, except for the static PES calculations of water molecules on the TiO_2 surface (Section 3), where a 2×2 distribution is employed. For systems which bear an electric dipole moment (e.g., the dry oxidized surface), we apply an electrostatic correction to remove the macroscopic dipole along the z supercell direction.[28] In all geometry relaxations we ensure that all force components on all unconstrained atoms are less than 0.05 eV/Å. Convergence of total energy differences with respect to the chosen cutoff is checked in all cases to be within 0.01 eV.

2.2. Classical MD Simulations. All our classical MD simulations are performed using the program package DLPOLY[29] (version 3.09), in which we have implemented the force field for the dry oxidized surface. The electrostatic interactions are treated using the smoothed particle mesh Ewald (SPME) method with a precision of 10^{-6} and a real space cutoff of 12.0 Å. If not stated otherwise, for the short-

ranged interactions a cutoff radius of 12.0 Å is used. Dynamic simulations at finite temperature are performed in a NVT ensemble using the Berendsen thermostat[30] with a relaxation time of 0.5 ps and an integration time step of 1 fs. The surface slab is constructed by repeating the DFT dry surface cell structure in each direction of the surface plane and by applying a mirror operation along the perpendicular direction in order to obtain a symmetric system without net dipole moment. The resulting surface areas of the supercells comprise 17.62 × 20.34 $Å^2$ for the 2 × 2 surface and 35.23 × 40.68 $Å^2$ for the 4 × 4 surface. In dynamic simulations of the oxidized surface and the TiO_2 slab, the surface atoms are allowed to move according to the force field described in ref 16 (which is also reported in detail in the Supporting Information, for completeness). In Sections 5.2 and 6, the central plane of titanium atoms is fixed to provide a constant reference coordinate frame. Before adding any adsorbate molecules, the dry surface is relaxed classically. The surface water interface is prepared by filling the vacuum gap with pre-equilibrated water molecules. After relaxing and thermalizing the system in a 200 ps MD run, the height of the simulation cell is initially adjusted in another 100 ps simulation to obtain a 1 atm pressure along the surface normal. Prior to each production run, further equilibration simulations of at least 200 ps are carried out. Adsorbate molecules are prepared by relaxing their structure in vacuum and placing them in the dry simulation cell, which is then filled with pre-equilibrated water molecules. Subsequently the system is treated as described above.

3. FPMD Simulations of Water Adsorption

In order to obtain accurate model systems of the interfaces between oxidized titanium and bulk water, we perform extended FPMD simulations based on DFT, comparing the water adsorption behavior on a rutile(110) surface with that on an ultrathin oxide film grown on Ti (0001).[16]

3.1. Water Adsorption on Rutile TiO_2 (110). The dominant adsorption mode of water on the rutile TiO_2 (110) crystal surface has been the subject of controversial discussions for decades. Theoretical studies have reported contradicting energetic orders for either molecular, mixed, or dissociative adsorption at low water coverage (≤1 ML).[18,31,32] Regarding experimental results, spontaneous dissociation of water molecules on the perfect rutile(110) surface is generally considered to be unlikely, whereas it is facilitated at surface defect sites.[33−35] Recently, the change of free energy, rather than of potential energy, upon water adsorption was calculated in DFT MD simulations, yielding a positive value of +0.6 eV for the dissociation of bulk water on the perfect rutile(110) surface,[36] which corroborates the experimental finding. Here we consider a 4-layer slab of a 1 × 3 surface unit cell including 24 titanium and 48 oxygen atoms in contact with 21 water molecules. The dimensions of our supercell are 6.56 × 8.95 × 40.0 $Å^3$. By means of both dynamical simulations and static total energy calculations, we find that molecular adsorption at the five-fold-coordinated titanium atoms (Ti_{5f}) is the preferred way of interaction on the perfect TiO_2 rutile surface, in agreement with the Car−Parrinello MD studies of ref 31. In particular, when starting from an initially dissociated configuration with one of the protons bound to the neighbor bridging oxygen atoms, proton transfer and recombination of the water molecule eventually occurs within a few hundred fs of dynamics. Direct Ti−O bond formation between water and the surface takes place exclusively at Ti_{5f} atoms with a coverage close to 100%. Namely, all three equivalent Ti_{5f} sites of our surface cell remain occupied by an adsorbed molecule for more than 90% of the time during the FPMD simulations at ∼350 K.

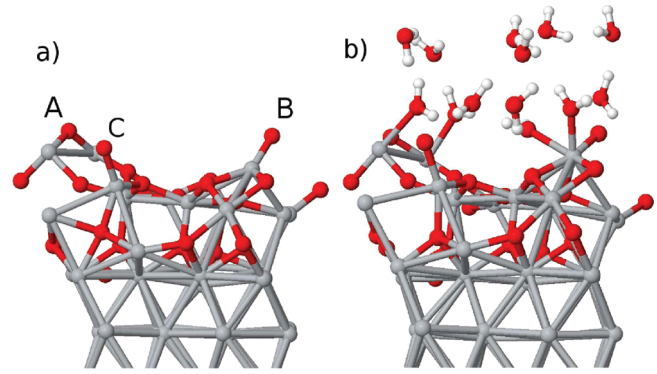


Figure 1. DFT model for the dry oxidized titanium surface (a) and snapshot of the interface between the surface and water from FPMD simulations (b).

3.2. Water Adsorption on Oxidized Ti(0001). A representative structural model for the oxidized Ti(0001) surface was obtained in extensive Car−Parrinello MD simulations, as described in ref 16 (Figure 1a). The model includes 60 titanium and 24 oxygen atoms corresponding to an O coverage of 2 ML. The structure of the oxide network exhibits a predominantly amorphous character, and its stoichiometry corresponds roughly to TiO, although features of different TiO_2 and Ti_2O_3 crystal structures can be identified.[16] As a remarkable topological property, the surface presents an exposed row of three two-fold-coordinated bridging oxygen atoms (labeled A, B, and C in Figure 1a), a typical feature observed on several TiO_2 crystal facets and on the oxidized TiN surface.[37]

Starting from this dry system we fill the vacuum gap with 28 pre-equilibrated water molecules and saturate the reactivity of the bottom surface of the slab with 12 hydrogen atoms in hcp positions, thus preventing spurious reactions between the water and the metallic slab. A FPMD simulation lasting 5 ps is carried out in the NVE ensemble, after initial thermalization of the system by velocity rescaling to a temperature of 350 K. During the dynamics, we observe adsorption, but not dissociation, of water molecules at exposed undercoordinated Ti atoms, similarly as on the rutile TiO_2(110) surface. In the case of the thin oxide film, the preferred adsorption sites are the Ti atoms which are bound to the two-fold-coordinated bridging oxygen atoms (which we will from now on refer to as TiB and OB, respectively). Once adsorbed, most of the molecules remain stably bound throughout the simulation. Only one molecule temporarily binds to a titanium atom located in the valley between the rows of bridging oxygen and later desorbs leaving the site free. In summary, a total of four water molecules stably adsorb on the surface during our FPMD trajectory, occupying three of the four TiB adsorption sites, one of which

accommodates two water molecules at the same time (Figure 1b). We calculate the desorption energies of these four water molecules from the total energy differences between the fully minimized water-decorated surface (in the absence of other free water molecules) and the same system upon removal of the adsorbed water molecules, one at a time, plus the total energy of the removed isolated water molecules in the same supercell. We obtain values of 0.53, 0.48, 0.91, and 0.44 eV. By comparison, for the desorption energy of a single water molecule from the fully hydrated 3 × 1 TiO_2(110) surface we find 0.83 eV.

In a further simulation, in which we started from a defective surface by removing one of the OB atoms, the adsorption of a water molecules takes place in a dissociative manner. One proton is transferred to a nearby bridging oxygen atom leaving a hydroxyl group adsorbed at the undercoordinated titanium atom. The corresponding calculated desorption energy value upon recombination of the dissociated molecule in the gas phase is 1.9 eV. Our computed values of desorption energy from the different sites agree fairly well with the values measured experimentally for the desorption of water from a ∼100 nm thick oxide layer on Ti.[38] Namely, two main desorption peaks at 0.53 and 0.75 eV were identified and assigned to desorption of molecularly adsorbed water on different surface sites, while a third peak at 1.2 eV was assigned to associative desorption from previously dissociated water molecules.

These results suggest that our model, although based on a system of very limited size, may indeed be representative of realistic Ti/TiO_x/water interfaces. We thus use it as a basis for constructing a classical potential which would enable us to simulate larger systems for longer time than achievable with a full quantum mechanical formalism. In doing this, we rely on the fact that our oxidized surface, in the absence of obvious defects, such as the oxygen vacancy that we arbitrarily created, showed little reactivity when exposed to liquid water. Therefore, we can assume that the physical/chemical behavior at the interface between oxidized Ti and the outer environment may be well captured by a simple potential based on nonbonded interactions, as described in the next section.

4. A Classical Potential for Ti/TiO_x/Water Interfaces

The starting point of our work is the classical force field potential which we have recently developed for the oxidized titanium surface. The model consists of a Finnis−Sinclair-type many-body potential for the metal region coupled to electrostatic Coulomb interactions and a short-ranged repulsive potential for the oxide region, as reported in the Supporting Information. In the next section, we focus on the coupling of this potential with a water environment and dissolved organic molecules, by combining it with standard molecular force fields.

4.1. Rescaling of the Point Charges for Surface/Adsorbate Interactions. In our potential, the point charges q_i employed in the electrostatic interactions within the oxide are determined using the electronegativity equalization method (EEM) of Mortier et al.[39] Their values within the thin oxide film (dashed line in Figure 2) are proportional to atomic Bader charges[40] computed at the DFT level and are consistent with the parametrization of the short-ranged interactions.[16] However, these charges are not guaranteed to reproduce well the electrostatic interactions between the oxide layer and the molecular species above the surface. Indeed, most common force fields for water or biomolecules, including the widely used TIP3P water model[41] or the AMBER[42] biomolecular force field, use point charges best fit to reproduce the electrostatic potential outside the molecule (ESP charges).[43−46] We thus compute ESP charges for the exposed OB and TiB atoms (for atoms buried in the surface, the ESP charge value has little or no significance). As shown in Figure 2, the obtained ESP charges are sightly lower than the EEM charges. Therefore, to compute the Coulomb electrostatic energy between molecular adsorbates and the oxidized surface, we rescale all surface point charges by a factor of 0.77, as determined ad hoc to match the EEM and ESP charges on the exposed surface bridging oxygen atoms (solid line in Figure 2). To calculate the interactions between the Ti and O atoms within the surface, we retain the original EEM charges in order to preserve the potential parametrization of ref 16.

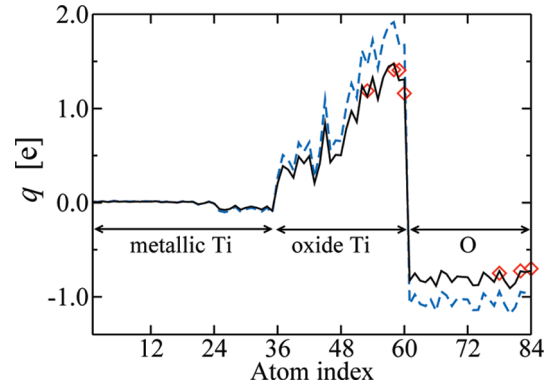


Figure 2. Charges of the dry oxidized titanium surface: Original EEM charges (dashed line, ---), scaled EEM charges (solid line, —), and the ESP charges of the exposed surface atoms (diamonds, ◊).

4.2. Interactions with Oxygen-Containing Molecules. The findings of Section 3, that water adsorption on the defect-free oxidized Ti surface takes place without dissociation, allow us to model the water/surface interactions by employing only electrostatic and nonbonded short-ranged forces. In this way, we can easily combine the potential described in the previous section with established biomolecular force fields in order to perform simulations of biomolecular adsorption on oxidized Ti, which is the ultimate goal of our work. We describe the interactions of the surface with adsorbates, in particular with water molecules, by a LJ and Coulomb nonbonded potential, as, e.g., in the AMBER force field:

$$V_{IJ}(r) = \varepsilon_{IJ}\left[\left(\frac{\sigma_{IJ}}{r}\right)^{12} - 2\left(\frac{\sigma_{IJ}}{r}\right)^{6}\right] + \frac{q_I q_J}{r} \qquad (1)$$

The parameters ε_{IJ} and σ_{IJ} for each pair IJ of interacting species can be obtained using the combination rules $\varepsilon_{IJ} = (\varepsilon_I \varepsilon_J)^{1/2}$ and $\sigma_{IJ} = (\sigma_I + \sigma_J)$.[42] The atomic parameters ε_I and

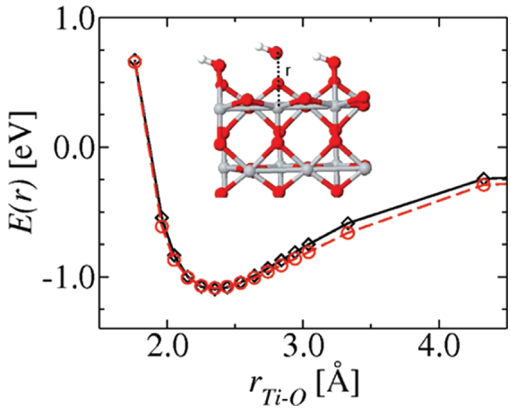


Figure 3. PES of a water molecule at various separations from the TiO_2 rutile(110) surface: DFT (black diamonds, ◊) and classical calculations (red circles, ○). The structure is displayed in the inset.

σ_I for water and biological molecules can be taken from the AMBER[42] or the generalized AMBER force field (GAFF),[47] which leaves only the four LJ parameters for the titanium and oxygen atoms of the surface to be determined. Differently from previous approaches (e.g., refs 20 and 21), we determine these parameters from a fit of the energy landscape of water desorbing from a rutile TiO_2(110) crystal surface, rather than from optimization of the structural properties of adsorbed water molecules.

As a reference, we compute with DFT the PES of one water molecule placed at different heights above the fully hydrated 3 × 1 rutile(110) surface cell, as displayed in the inset of Figure 3. Starting from a fully minimized configuration, one of the water molecules is displaced vertically along the surface normal, and total energy calculations are performed at each separation, keeping all atomic positions fixed. The resulting PES is shown in Figure 3 (black solid line), yielding a potential minimum of −1.1 eV. This is deeper than the desorption energy computed in Section 3 (0.83 eV) because of the lack of atomic relaxation.

For exactly the same atomic configurations, we now compute total energies using our classical potential, optimizing the LJ parameters for the surface atoms by a least-squares fit to the DFT PES using the GULP package.[48] Both the DFT and the classical energy values are rigidly shifted to obtain a value of 0.0 eV for a water surface separation of 8 Å. In these calculations, the point charges on the rutile atoms are computed from the EEM charges scaled by the same factor of 0.77 determined for the oxidized Ti surface (see previous section). Notably, the resulting charges are nearly identical to the ESP charges computed for the crystal surface (e.g., the average charge values of the bridging oxygen atoms are 0.67 electrons in both cases).

As shown in Figure 3, the agreement between the DFT and classical PES is excellent for the optimal LJ parameter set listed in Table 1. We note that in our approach the used LJ potential must not be seen as a physical representation of dispersion interactions but only as an arbitrary way of mapping the true surface water interactions by means of Coulomb and short-range terms. In fact, weak dispersion interactions are not properly accounted for, and generally underestimated,[49] in the DFT total energy calculations.

Table 1. LJ Parameters of the Surface Atoms

	LJ parameters	
	ε_I [eV]	σ_I [Å]
Ti	0.01455	0.7827
O	0.01983	1.6154
	Ti−N 9−6 potential	
ε_{Ti-N} [eV]	0.140155	
σ_{Ti-N} [Å]	2.30769	

Table 2. Interatomic Distances of the DFT and the Classical Model after Relaxation of the Water Molecules on the Rutile(110) Surface

	DFT	classical
Ti_5−OW [Å]	2.34	2.24
HW−OB [Å]	1.81	2.09
HW−OW [Å]	2.14	2.17

However, the deep minimum of the potential well on the polar oxide surface suggests that electrostatic attraction by far exceeds the dispersion forces, which can be thus safely neglected.

Using these interaction parameters, the water molecules are relaxed classically to compare the adsorbed geometry on the crystal surface to the corresponding DFT structure. Upon full atomic relaxation, at the classical level, the calculated desorption energies of a single water molecule from the hydrated surface is 0.81 eV, which is in very good agreement with the DFT value of 0.83 eV. We will indeed show later in the paper that these parameters lead to computed values of the work of hydration of Ti oxide surfaces in good agreement with experiments, thus justifying the approximations taken in our approach. The distances between five-fold titanium and water oxygen, between water hydrogen and bridging oxygen, as well as between hydrogen and oxygen of two neighbor water molecules are reported in Table 2. We notice small differences between the DFT and the classical structure, in particular the hydrogen bridges are longer in the classical model. However, we consider these differences as acceptable for our purposes, and we refrain from correcting them ad hoc by introducing bending potentials,[20,21] as they would prevent the desorption of the bound water molecules from the surface, or their replacement by other water molecules. These are events that we often observed in long FPMD simulations and that we would like to reproduce also in classical simulations.

Another feature which should be captured by the potential is the correct adsorption energy of a second water layer, as discussed in ref 21. To check this issue, we place an additional water molecule over the crystal surface terminated by three adsorbed water molecules, with the H atoms pointing toward the surface OB atoms. For this system we calculate the DFT and the classical PES as described above, obtaining adsorption energy minima of −0.14 and −0.12 eV for the DFT and the classical potential, respectively.

4.3. Interactions with Nitrogen-Containing Molecules. Obviously not all molecules of interest bind to the surface via oxygen atoms, as in the case of water. It is thus necessary

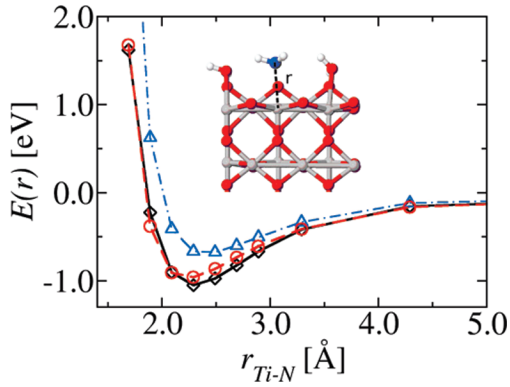


Figure 4. PES of an ammonia molecule at various separations from the TiO_2 rutile 110 surface: DFT (black diamonds, ◊) compared to classical calculations with original (blue triangles, △) and modified parameters (red circles, ○). The structure is displayed in the inset.

to check the transferability of the surface LJ parameters to the case of molecules adsorbing via different atoms, in particular nitrogen given its abundance in protein and other biomolecules. For this purpose we compute the adsorption of ammonia on the hydrated TiO_2(110) surface, as the simplest possible reference case. PES calculations are performed in the same way as described above, replacing only one of the water molecules by ammonia, while retaining the other two water molecules adsorbed on the surface. For the classical description of the NH_3 molecule, we use the LJ parameters taken from the GAFF. Since the AMBER force field does not specify partial charges for ammonia, we assign to the atoms ESP charges of −0.84 and +0.28, obtained by the same method as described in Section 4.1. Comparing the DFT and classical PES (Figure 4), it appears that the position of the energy minimum determining the equilibrium bond length is slightly shifted toward larger distances, and especially the depth of the potential minimum is too shallow in the classical case. Notably, even trying a further optimization of the LJ parameter of the surface Ti and O atom did not lead to satisfactory results. Depending on the particular circumstances, this deviation from the quantum mechanical behavior can be accepted either as a limit of the parameter transferability or a modification of the potential form must be introduced. In order to ensure a tight consistency with the DFT results, we chose to introduce an ad hoc 9-6 potential to model the interactions between N and Ti atoms:

$$V_{\mathrm{Ti-N}}(r) = \frac{\varepsilon_{\mathrm{Ti-N}}}{3}\left[6\left(\frac{\sigma_{\mathrm{Ti-N}}}{r}\right)^9 - 9\left(\frac{\sigma_{\mathrm{Ti-N}}}{r}\right)^6\right] \qquad (2)$$

The parameters $\varepsilon_{\mathrm{Ti-N}}$ and $\sigma_{\mathrm{Ti-N}}$ are determined by fitting to the DFT PES, the resulting values given in Table 1. With this potential form, the DFT adsorption energy profile can be now very well reproduced (Figure 4).

4.4. Adsorption of Organic Molecules on the Dry Oxidized Ti Surface. In this section, we check whether the force field parameters determined in the previous section taking the TiO_2(110) surface as a reference are transferable to the case of adsorption of small organic molecules on the oxidized Ti surface. To this aim, the PES of methanol (CH_3OH), formic acid (HCOOH), and methylamine (CH_3NH_2) above the dry oxidized titanium surface are calculated both by means of full-level DFT and of our newly developed classical potential. For the reasons mentioned in Section 4.3 and for the sake of consistency, for all molecules we computed ESP charges with our DFT code. These are found to differ by less than 0.05 e from the corresponding point charges of the AMBER force field, when available. The LJ parameters of all atomic pairs are obtained by the standard combination rules, as described above. For each of the molecules, the minimum-energy adsorption geometry is obtained by FPMD simulations followed by careful relaxation. Taking the resulting structures as the input models, the molecules are displaced along the directions of the bond connecting them to the surface, and total energy calculations are performed without atomic relaxation. The relaxed adsorbed configurations are shown in Figure 5.

In the case of formic acid, we found that the molecule could adsorb in either a molecular or a dissociated form, depending on the initial orientation of the carboxyl hydrogen. Since the dissociation reactions cannot be taken into account using our simple force field, we focus here on the molecularly adsorbed configuration. The OH group of methanol was found to bind to two titanium atoms, therefore the molecule was displaced vertically above the surface. Methylamine

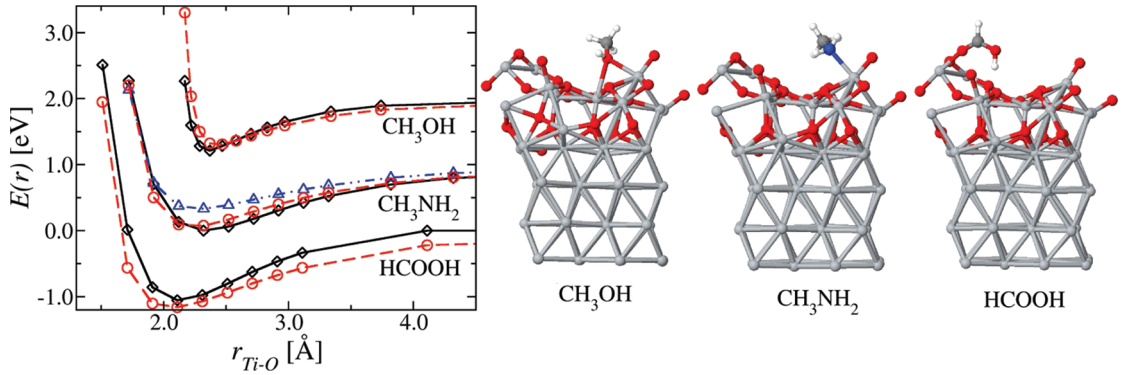


Figure 5. PES of methanol (CH_3OH), methylamine (CH_3NH_2) and formic acid (HCOOH) on the dry oxidized titanium surface: DFT (black diamonds, ◊) vs classical energies (red circles, ○). For CH_3NH_2 the results for unchanged (blue triangles, △) and modified (red circles) Ti−N interactions are displayed. For clarity the PES for CH_3NH_2 and CH_3OH are shifted vertically by 1.0, respectively, 2.0 eV.

adsorbs with the N atom of the amine group bound to a TiB atom of the surface.

For methanol, the PES obtained with our classical potential agrees very well with the energies calculated with DFT (Figure 5), in the case of formic acid the classical energies slightly overestimate the DFT values by about 0.1 eV. For methylamine, we find excellent agreement between the two PES when including the modified 9-6 Ti−N interaction potential, whereas using the standard 12-6 LJ potential for Ti−N results in considerably lower adsorption energy, although the equilibrium bond length is correctly reproduced.

5. Adsorption Behavior of Wet Systems

In the previous section, we have constructed a classical force field potential which is able to reproduce the adsorption energy of small molecules on both crystalline TiO_2 surfaces and thin oxide films grown on Ti(0001). Here, we apply our potential to investigate the behavior of interfaces between oxidized Ti and liquid water or fully solvated organic molecules. In particular, we take into account two representative cases for which quantitative experimental results are available, namely the heat of immersion of titanium oxide and the adsorption of single tyrosine molecules on oxidized Ti.

5.1. Heat of Immersion of TiO_x Surfaces. The heat of immersion of a surface, ΔH_{imm}, is defined as the energy gained upon placing the dry surface in contact with liquid water. In contrast to the case of, e.g., the oxidized silicon surface, where water molecules are stably chemisorbed in a dissociative manner, mostly molecular adsorption and physisorption of water takes place on titanium oxide surfaces, as already mentioned in Section 3. As found in TPD experiments, the desorption temperature of these molecules is around or even below room temperature.[34,38,50] Therefore, the amount of surface water molecules which remain bound to the surface upon drying cannot be unambiguously identified, as this significantly depends on the conditions of preparation, in particular on the drying temperature.[50] Correspondingly, as ΔH_{imm} depends on the number of molecules already bound to the surface prior to immersion in liquid water, scattered values between 0.2 and 0.6 J/m^2 have been reported for TiO_2 crystals.[50−52] A linear decrease of ΔH_{imm} with an increasing amount on initially adsorbed water on the TiO_2 rutile and anatase surfaces has been obtained in ref 50. Also in this study, however, the measured values scatter by as much as 0.3 J/m^2 for different investigated samples at the same initial water coverage, which makes possible only a rough comparison with theoretical investigations.

Here we start our study considering the interface between bulk water and a six-layer slab model of the rutile TiO_2(110) surface including a 6 × 12 surface unit cell comprising an area of 35.23 × 40.68 Å^2. Similar to what is observed in FPMD simulations, in classical MD runs at 300 K we observe water molecules binding preferentially to Ti_{5f} atoms, where they remained stably adsorbed for large part of the simulations. The heat of immersion can be calculated by subtracting from the average potential energy of the wet surface E_W the average potential energy of the corresponding dry surface E_D and the potential energy of bulk water containing the remaining number of water molecules E_B.[53] Starting from a 200 ps trajectory of the whole system with N_{tot} water molecules, these molecules are sorted with decreasing probability of being bound to the surface. The bulk water molecules are removed leaving only a certain number N_{ads} of molecules (according to their adsorption probability) on the surface. For these dry surfaces, simulation runs of 200 ps at a temperature of 300 K are performed to obtain the corresponding E_D average potential energies. From these values and the corresponding potential energy of bulk water containing $N_{tot} - N_{ads}$ water molecules, we calculate the heat of immersion as

$$\Delta H_{imm}(N_{ads}) = [E_D(N_{ads}) + E_B(N_{tot} - N_{ads}) - E_W(N_{tot})]/(2 * A_{Surf}) \quad (3)$$

where A_{Surf} is the surface area of only one side of the slab. The potential energies of the bulk water systems are calculated by first adjusting the height of each water cell in a 200 ps NPT run to obtain a pressure of 1 atm, followed by another 200 ps NVT simulation, in which the average energies were computed. A total number of 1188 water molecules is included as the liquid phase, and 'dry' surface systems with 24, 36, 64, 88, 120, and 144 preadsorbed water molecules are investigated. For N_{ads} = 144, all five-fold coordinated Ti atoms of the rutile 110 surface are occupied by water molecules.

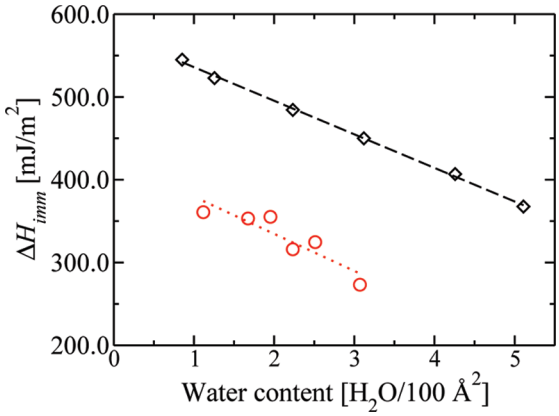


Figure 6. Heat of immersion for the TiO_2 rutile 110 surface (black diamonds, ◊) and the oxidized Ti surface (red circles, ○) as a function of the water content. The straight lines are linear fits to the data.

The resulting dependence of ΔH_{imm} on the number of preadsorbed water molecules is shown in Figure 6. In agreement with the findings of ref 50, a perfectly linear decrease is obtained, and also the absolute values compare well with those available in the literature (between 200 and 600 mJ/m^2, see above). The slope of the linear regression is 0.25 eV/H_2O, which represents the desorption energy per molecule from the surface into bulk water. If we neglect the hydration of adsorbed molecules, then the same quantity can be calculated by adding the heat of vaporization of water (−0.45 eV for TIP3P water)[41] to the desorption energy into the gas phase (∼0.8 eV, see above), obtaining 0.35 eV. A

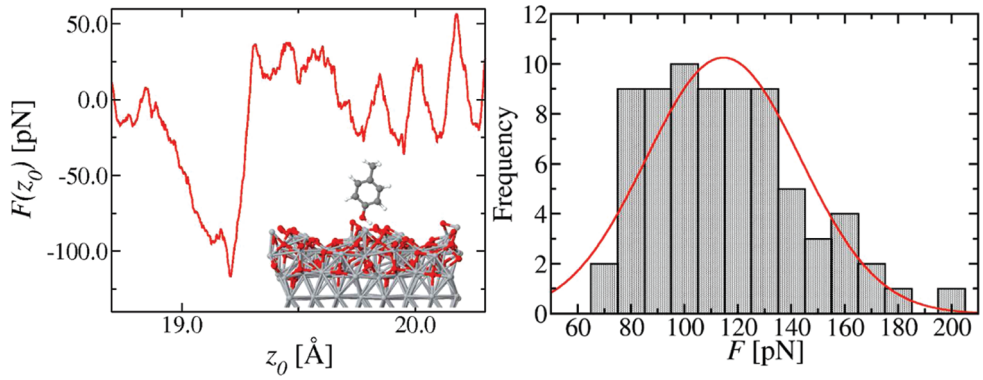


Figure 7. SMD simulations of tyrosine on the oxidized surface. Left: Example for a force−displacement curve $F(z_0)$. Right: Histogram of the maximum desorption forces and Gaussian fit to the distribution (red line).

comparison of these two numbers gives an estimate of about −0.1 eV for the hydration energy of adsorbed molecules on the surface.

We note that our value of 0.25 eV only takes into account molecular adsorption, while dissociative adsorption events on defective crystal sites, such as steps or edges, are expected to be associated with larger energy values (of the order of 1.2 to 1.9 eV, see Section 3). This may explain the significantly larger value of 0.82 eV/H_2O reported in ref 50 for the case of rutile powder samples.

The heat of immersion calculated for the oxidized surface displays a slightly different behavior. In this case, we include 1520 water molecules in the liquid phase in contact with the 4 × 4 repetition of the DFT surface model and performed 'dry' simulations for N_{ads} = 32, 48, 56, 64, 72, and 88. First of all, the obtained values are considerably lower than those for the crystal surface, as they range from about 260 to 380 mJ/m^2. They are closer to the value of 260 mJ/m^2 which has been reported for small TiO_2 nanoparticles.[51] Moreover, although a tendency of the heat of immersion to decrease upon increasing the preadsorbed water content can be identified, the values are scattered, and no clearly linear dependence is observed. This must be attributed to the fact that, in contrast to the perfect rutile surface, not all adsorption sites are equivalent on the oxidized surface, as indicated also by the scattered values of the static DFT adsorption energies on this surface (cf., Section 3). The scattering could possibly be reduced by changing the order of removing the water molecules and thus averaging over the different adsorption sites. However, sampling of a large number of permutations would increase the computer time exceedingly and lies beyond the scope of this work. Interestingly, fitting a linear function to the obtained values yields to a very similar slope compared to the crystal surface.

5.2. Desorption Force of Tyrosine. As a further validation of the force field, we calculate the maximum detachment force of single tyrosine residues from the oxidized titanium surface. This has been measured by AFM force spectroscopy experiments leading to a value of 97 ± 28 pN.[54] In our simulations, to exclude contributions from the backbone adsorbing to the surface, we consider a reduced molecule consisting of a phenol ring bound to a methyl group. The intramolecular interactions as well as the LJ parameters and the partial charges of the molecule are taken from the AMBER force field, and the charge value of the methyl carbon was adjusted to obtain a neutral molecule. After equilibration we then carry out a 39 ns classical MD run of the molecule on the oxidized surface at 300K, recording one snapshot every 500 ps. Seventy-eight of these snapshots were taken as independent starting configurations in subsequent umbrella-sampling runs. Such a large number of simulations yields reliable statistics for the force distribution, however, as a drawback, only the small 2 × 2 repetition of the DFT surface model could be used, to keep the computational cost reasonable. Due to the smaller cell size the cutoff radius for nonbonded interactions and for the real-space contribution of the electrostatic interactions had to be reduced to 8.0 Å.

Using a harmonic umbrella potential in the z direction normal to the surface applied to the carbon atom of the methyl group:

$$V_{umbr}(z_c) = \frac{1}{2}k_{umbr}(z_c - z_{umbr})^2 \quad (4)$$

with k_{umbr} = 0.2 eV/Å^2 and z_{umbr} = 16.0 Å (compared to $z \simeq 12$ Å for the exposed bridging oxygen atoms of the surface), the molecule was initially constrained to be close to the surface in a 300.0 ps simulation. In this representation, the z = 0.0 value refers to the central plane of titanium atoms, which are kept fixed. A steered molecular dynamics simulation (SMD) was then performed to mimic the experimental AFM setup, applying a time-dependent umbrella potential:

$$V_{smd}(z_c, t) = \frac{1}{2}k_{smd}(z_c - z_0(t))^2 \quad (5)$$

By choosing $z_0(t) = z_c(t = 0) + v_{smd} \cdot t$, the molecule is pulled off the surface at constant velocity. We set v_{smd} = 0.5 m/s and k_{smd} = 0.1 eV/Å^2. The instant pulling force $F(z_0) = k_{smd}(z_c - z_0)$ is recorded as a function of the pulling height z_0 every 5 fs. In order to eliminate large fluctuations, running averages of the force values over blocks comprising z_0 ranges of 0.025 Å are taken into account. In this way the short-time fluctuations are found to decrease considerably, whereas the actual force−displacement curve, which varies on a larger time scale, is not affected significantly.

A representative example of a force−displacement curve is shown in Figure 7. Initially the adhesion force increases roughly linearly until eventually a sudden decrease is visible,

which reflects the detachment of the molecule from the surface. We calculate the peak forces for a total number of 73 simulations, their distribution is displayed in a histogram in Figure 7. In five cases no clear force peak could be identified, indicating that the molecule was not adsorbed at the beginning of the simulations (more precisely, we did not consider peaks smaller than 60 pN, which corresponds to the fluctuations of the pulling force acting on a free, solvated molecule dragged through bulk water). These simulations were discarded and not considered in the histogram. Considering the trajectories of the individual simulations, we note that the adhesion to the surface is in general mediated by hydrogen bonds between the hydroxyl group of the phenol ring and the surface oxygen atoms, as assumed in ref 54. Moreover, in some cases the hydroxyl oxygen is observed to bind temporarily to one TiB atom after displacement of an adsorbed water molecule, leading to the values on the shoulder toward larger forces in the distribution. In summary, our computed forces range from 70 to 200 pN. A Gaussian distribution fit to the values yields an average force of 108 pN and a width of 31 pN. Within this variance the average force value agrees well with the experimental results of ref 54 (97 ± 28 pN). Therefore, we feel that our interaction potential can be reliably applied to investigate new systems, for which the experimental understanding is still incomplete, as performed in a representative case in the next section.

6. Adsorption of RGD Peptides on the Oxidized Ti Surface

Finally, as a first application of the developed force field, we present simulations of the RGD peptide sequence adsorbing on Ti. This sequence is present in proteins building the extracellular matrix, such as fibronectin and collagen, where it acts as an integrin binding site and plays an important role in the process of cell adhesion.[55] Since such peptides are used to functionalize the surfaces of metal implants to enhance bone cell adhesion,[56,57] an interesting aspect is their direct adsorption behavior, as this competes to binding to integrins. Despite its importance, only a few simulation studies are devoted to the investigation of the adsorption of RGD sequences on solid-state surfaces, particularly on crystalline titanium oxide.[14,15,58]

Here we perform umbrella sampling simulations to obtain force−displacement profiles from which the potential of mean force (PMF) and the free energy of adsorption can be calculated. In order to avoid charged end groups, the molecule is terminated with NME (CH_3NH-) and ACE ($-COCH_3$) sequences, yielding a NME-Asp-Gly-Arg-ACE peptide. The peptide is completely modeled using the AMBER force field, including its partial charges. For similar reasons as stated in Section 5.2, we consider a 2 × 2 surface area and use a cutoff radius of 8.0 Å. After pre-equilibrating and adjusting the cell height, the system is annealed at 450 K for 200 ps (keeping the surface atoms fixed) to overcome possible adsorption barriers to the surface, followed by another annealing at 300 K for 300 ps. The resulting configuration, which is shown in Figure 8c, is used as initial model for our free energy calculations.

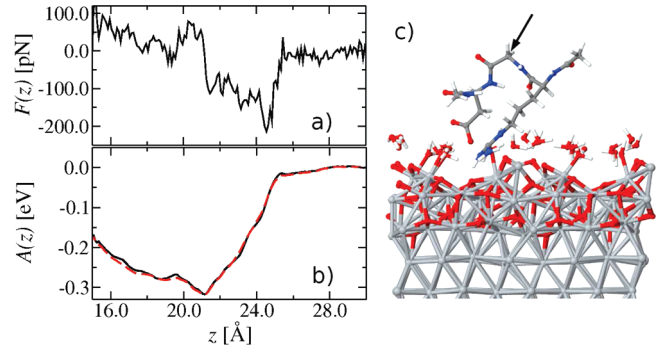


Figure 8. Desorption of the RGD-containing peptide from the oxidized Ti surface: Force profile (a), free energy profile obtained by WHAM (solid line) and TI (dashed line), and snapshot of the initial adsorbed configuration (c). For clarity, only the first layer of water molecules is displayed. The arrow marks the Gly α-carbon atom.

The RGD molecule binds to the surface via the Arg side chain, which is able to penetrate the first layer of water molecules in the valley between two rows of bridging oxygen atoms. The interaction with the surface is mediated both via hydrogen bonds between the guanidine group and surface oxygen atoms and via electrostatic interactions between nitrogen and titanium atoms. The ASP side chain is also oriented toward the surface, with the carboxyl oxygen atoms forming hydrogen bonds with the Arg side chain and with surface water molecules.

In the umbrella sampling simulations, as the reaction coordinate we chose the z-position $z_{c\alpha}$ of the α carbon of the central Gly residue, with a zero offset corresponding to the position of the central plane of the surface slab (as described in Section 5.2). The reaction coordinate is then increased stepwise from 16.0 to 29.0 Å and restrained to a total of 14 windows with 1.0 Å width by a harmonic potential (see Section 4) with a force constant $k_{umbr} = 0.2$ eV/Å^2. For each window a simulation run of 1.2 ns is performed, where the first 200 ps are discarded from the force analysis. The reaction coordinate and the z-component of the force acting on it are recorded every 5 fs.

To calculate the free energy profile $A(z)$, we employ two conceptually different methods, namely: (i) the weighted histogram analysis method (WHAM),[59] evaluating the probability using the code of Grossfield,[60] and (ii) the PMF as obtained by thermodynamic integration (TI) of the average force:[61,62]

$$A(z) = -\int_{z_{max}}^{z} \langle F(z')\rangle dz' \qquad (6)$$

The unbiased force is obtained by performing the average over all umbrella windows and correcting the value by the respective umbrella force:

$$\langle F(z)\rangle = \sum_{i=1}^{N_{umbr}} \frac{n_i(z)[\langle F^i_{biased}(z)\rangle + (dV^i_{umbr}(z)/dz)]}{n_{tot}(z)} \qquad (7)$$

where the i indicates the respective umbrella window, $n_i(z)$ gives the number of appearances of a reaction coordinate value of z from umbrella window i, and $n_{tot}(z) = \sum_i n_i(z)$

yields the total number of counts for the value z. The forces, probability, and corresponding free energy profile are collected in bins of 0.1 Å width.

The unbiased force profile $F(z)$ is displayed in Figure 8a. When increasing the z_0 value of the umbrella center, first the side chain of the Asp residue is detached from the surface due to its shorter length compared to Arg. Finally the guanidine group of the Arg residue desorbs producing a force peak of about −215 pN at a z-value of 24.5 Å in the force profile. The free energy profiles calculated with the two methods are shown in Figure 8b. Importantly, we note that the two curves agree almost perfectly with each other, giving a strong hint that the force calculations and the reaction coordinate sampling have reached convergence. In the free energy profile we observe a minimum depth of 0.32 eV, which can be interpreted as the free energy of desorption. Experimental values for the binding free energy between RGD-containing peptides and integrin proteins in the absence of a surface are found to be in the range of 0.16 eV,[63] whereas simulations of such a situation yield a binding free energy of 0.13 eV.[64] Hence, when a titanium surface is functionalized using RGD-containing peptides, a situation might arise where adsorption on the surface is in competition with the desired process of binding to integrin molecules. From a comparison of the respective free energy values we can conclude that the adsorption of RGD on the oxidized titanium surface is considerably stronger and might thus limit the functionality of the sequence. Therefore, covalent biding via appropriate spacers, where direct adsorption of RGD at the titanium surface is avoided, should be preferred over nonspecific surface adsorption to enhance cell adhesion via binding to integrin.

7. Conclusions

In summary, we have presented an extension of the classical force field developed in ref 16 to model the interactions between natively oxidized titanium surfaces and liquid water as well as solvated biomolecules. The interactions across the solid/liquid interfaces comprise Coulomb forces between ESP point charges and a L J potential, whose coefficients for the surface atoms have been determined by fitting the classical PES of a water molecule at various separations from the TiO_2 rutile 110 surface to the corresponding DFT energies. In this way, the potential is fully consistent with commonly used biomolecular force fields. We have demonstrated that the interactions with generic organic molecules can be reliably obtained by applying standard combination rules to the GAFF. In particular, the obtained potential is fully transferable to the case of molecules containing O, C, and H atoms adsorbed on thin oxide layers grown on metallic Ti, for which the adsorption PES calculated with full DFT and with our classical potential is excellent. However, if the direct surface-molecule interactions involve nitrogen atoms, quantitative agreement between the DFT and classical PES could be obtained only after introducing an additional 9-6 potential to model the Ti−N interactions. After adjusting the respective potential parameters, using an NH_3 molecule adsorbed on the partially wet rutile surface as a reference, excellent transferability to the case of the natively oxidized surface has been found.

As mentioned before, the major approximation intrinsic in our potential parametrization is the use of standard DFT calculations to determine the reference surface/molecule interactions, which do not properly take into account dispersion forces. If necessary, more sophisticated methods to compute the reference PES should be employed, and the LJ parameters of the interaction potentials correspondingly adjusted, while the chosen analytic form of the potential certainly allows for dispersion forces to be described correctly. However, in our specific case where highly polar surfaces are considered, the electrostatic contributions far exceed weak forces of the van der Waals type, resulting in adsorption energies of the order of 0.8 eV per water molecule. Indeed, with our potential parametrization we obtain a fairly good agreement between the absolute values of the computed and the measured heat of immersion of TiO_2 crystals as well as of the maximum adhesion force of single Tyr molecules to Ti surfaces. The latter has been obtained by means of steered MD simulations, using a time-dependent harmonic spring potential to pull the Tyr side chain off the surface. The average of the force peaks, 108 ± 31 pN, is in good agreement with the only available measured value of 97 ± 28 pN, and the computed and measured distributions present a similar standard deviation.

As far as the heat of immersion of TiO_2 is concerned, we have found a linear decrease with increasing water content chemisorbed on the surface prior to immersion in liquid water. This is consistent with the experimental study of ref 50, in which the reported range of energy values agrees very well with the simulation results. However, as already mentioned, the slope of $\Delta H_{imm}(N_{ads})$ is significantly smaller than the correspondent experimental value, probably because of dissociative adsorption events which may take place over the surface of powder crystals samples but cannot be explicitly taken into account in our classical model.

In fact, a potential as simple as the one presented here (based on purely electrostatic and LJ interactions) is expected to be accurate only under the assumption that no bond breaking or forming events take place, except the direct binding of O or N atoms of organic molecules to Ti atoms of the surface, for which the potential has been parametrized ad hoc. Under this assumption, the transferability of our potential to the case of generic organic molecules on the oxidized titanium surface is surprisingly good and allows us for the first time to investigate the atomistic mechanisms of biomolecular adsorption at titanium/water interfaces.

As a preliminary example, we have studied the adsorption of solvated RGD tripeptides on the oxidized Ti(0001) surface. Considering one possible adsorption mode, where the Arg side chain adsorbs at the surface via hydrogen bonds, we have found a free energy of desorption of 0.32 eV. The corresponding maximum detachment force reaches a value of 215 pN. As mentioned in previous publications,[14,15,58] several adsorption modes involving different side chains are possible on titanium oxide surfaces. In the context of this investigation we have restricted ourselves to just one initial configuration, in order to demonstrate the applicability of

the force field. A thorough investigation of the adsorption behavior and the corresponding free energies of RGD-containing peptides will be the subject of future work.

Future work will also be concerned with the application of the potential to larger biomolecules, which are relevant in cell adhesion processes. As already pointed out in this work, one of the main challenges in this kind of simulation will be the calculation of adsorption free energies, which becomes increasingly difficult for more complex systems. Furthermore, a possible extension of the model would be to take into account surface defects and dissociative water adsorption. However, we feel that the best way to proceed further in this direction is to implement our simple force field in hybrid QM/MM simulation schemes, such as, e.g., the Learn on the fly (LOTF) method.[65] This would enable a quantum mechanical treatment of the chemically active system regions, e.g., at the solid/liquid interface, while allowing at the same time the inclusion of a realistically large model of the physiological environment.

Acknowledgment. We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) under the Emmy Noether grant CI 144/2-1 and from the EU-FP7-NMP grant 229205 "ADGLASS". Computer time was allocated at the HLRN (Hannover-Berlin) and the ZIH (Dresden) supercomputing centers.

Supporting Information Available: Structures and charges of the oxidized titanium surface and of the TiO_2 rutile 110 surface as well as the partial charges for the adsorbate molecules (if not taken from the AMBER force field) are available. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Takemoto, S.; Hattori, M.; Yoshinari, M.; Kawada, E.; Oda, Y. *Biomaterials* **2005**, *26*, 6014–6023.

(2) Aziz-Kerrzo, M.; Konroy, K. G.; Fenelon, A. M.; Farrell, S. T.; Breslin, C. B. *Biomaterials* **2001**, *22*, 1531–1539.

(3) Horbett, T. A.; Brash, J. L. In *Proteins at Interfaces II*; Horbett, T. A., Brash, J. L., Eds.; ACS Symposium Series: Washington, DC, 1995; Chapter 1, pp 1−23.

(4) Oviedo, C. *J. Phys.: Condens. Matter* **1993**, *5*, 153–154.

(5) Takakuwa, Y.; Ishidzuka, S.; Yoshigoe, A.; Teraoka, Y.; Yamamauchi, Y.; Minzuno, Y.; Tonda, H.; Homma, T. *Appl. Surf. Sci.* **2003**, *216*, 395–401.

(6) Burrell, M. C. *J. Vac. Sci. Technol., A* **1983**, *1*, 1831–1936.

(7) Liu, S.-Y.; Wang, F.-H.; Yun-Song-Zhou; Shang, J.-X. *J. Phys.: Condens. Matter* **2007**, *19*, 226004–226015.

(8) Azoulay, A.; Shamir, N.; Fromm, E.; Mintz, M. H. *Surf. Sci.* **1997**, *370*, 1–16.

(9) Vaquila, I.; Passeggi, M. C. G.; Ferron, J. *Appl. Surf. Sci.* **1996**, *93*, 247–253.

(10) Köppen, S.; Ohler, B.; Langel, W. *Z. Phys. Chem.* **2006**, *221*, 3–20.

(11) Köppen, S.; Langel, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1907–1915.

(12) Carravetta, V.; Monti, S. *J. Phys. Chem. B* **2006**, *110*, 6160–6169.

(13) Skelton, A. A.; Liang, T.; Walsh, T. *ACS Appl. Mater. Interfaces* **2009**, *1*, 1482–1491.

(14) Liang, Y.-C.; Song, D.-P.; Chen, M.-J.; Bai, Q.-S. *J. Vac. Sci. Technol., B* **2009**, *27*, 1548–1554.

(15) Wu, C.; Chen, M.; Guo, C.; Zhao, X.; Yuan, C. *J. Phys. Chem. B* **2010**, *114*, 4692–4701.

(16) Schneider, J.; Colombi Ciacchi, L. *Surf. Sci.* **2010**, *604*, 1105–1115.

(17) Li, W.-K.; Lu, G.; Selloni, A. *J. Phys. Chem. C* **2008**, *112*, 6594–6596.

(18) Lindan, P. J. D.; Zhang, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *72*, 075439−075445.

(19) Köppen, S.; Langel, W. *Surf. Sci.* **2006**, *600*, 2040–2050.

(20) Bandura, A. V.; Kubicki, J. D. *J. Phys. Chem. B* **2003**, *107*, 11072–11081.

(21) Skelton, A. A.; Walsh, T. *Mol. Simul.* **2007**, *33*, 379–389.

(22) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244–13249.

(23) Blöchl, P. E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953–17979.

(24) Vita, A. D.; Canning, A.; Galli, G.; Gygi, F.; Mauri, F.; Car, R. *EPFL Supercomput. Rev.* **1994**, *6*, 22.

(25) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(26) VandeVondele, J.; Vita, A. D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *60*, 13241–13244.

(27) Stengel, M.; Vita, A. D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *62*, 15283–15286.

(28) Neugebauer, J.; Scheffler, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 16067–16080.

(29) Todorov, I. T.; Smith, W. *Phil. Trans. R. Soc., A* **2004**, *362*, 1835–1852.

(30) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(31) Langel, W. *Surf. Sci.* **2002**, *496*, 141–150.

(32) Barnard, A. S.; Zapol, P.; Curtiss, L. A. *Surf. Sci.* **2005**, *582*, 173–188.

(33) Diebold, U. *Surf. Sci. Rep.* **2003**, *48*, 53–229.

(34) Henderson, M. A. *Surf. Sci.* **1996**, *355*, 151–166.

(35) Wendt, S.; Matthiesen, J.; Schaub, R.; Vestergaard, E. K.; Lægsgaard, E.; Besenbacher, F.; Hammer, B. *Phys. Rev. Lett.* **2006**, *96*, 066107−066110.

(36) Cheng, J.; Sprik, M. *J. Chem. Theory Comput.* **2010**, *6*, 880–889.

(37) Zimmermann, J.; Finnis, M. W.; Colombi Ciacchi, L. *J. Chem. Phys.* **2009**, *130*, 134714–134725.

(38) Lausmaa, J.; Loefgren, P.; Kasemo, B. *J. Biomed. Mater. Res.* **1999**, *44*, 227–242.

(39) Mortier, W. J.; Genechten, K. V.; Gasteiger, J. *J. Am. Chem. Soc.* **1985**, *107*, 829–835.

(40) Bader, R. F. W. In *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, U.K., 1994; Chapter 6, pp 169−247.

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

(42) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.;

Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(43) Momany, F. A. *J. Phys. Chem.* **1978**, *82*, 592–601.

(44) Cox, S. R.; Williams, D. E. *J. Comput. Chem.* **1981**, *2*, 304–323.

(45) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.

(46) Woods, R. J.; Khalil, M.; Pell, W.; Moffat, S. H.; Smith, V. H., Jr. *J. Comput. Chem.* **1990**, *11*, 297–310.

(47) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(48) Gale, J. D. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 629–637.

(49) Ortmann, F.; Schmidt, W. G.; Bechstedt, F. *Phys. Rev. Lett.* **2005**, *95*, 186101–186104.

(50) Morimoto, T.; Nagao, M.; Omori, T. *Bull. Chem. Soc. Jpn.* **1969**, *42*, 943–946.

(51) Gun'ko, V. M.; Blitz, J. P.; Zarko, V. I.; Turov, V. V.; Pakhlov, E. M.; Oranska, O. I.; Goncharuk, E. V.; Gornikov, Y. I.; Sergeev, V. S.; Kulik, T. V.; Palyanytsya, B. B.; Samala, R. K. *J. Colloid Interface Sci.* **2009**, *330*, 125–137.

(52) Dawber, J. G.; Guest, L. B.; Lambourne, R. *Thermochim. Acta* **1972**, *4*, 471–484.

(53) Cole, D. J.; Csanyi, G.; Payne, M. C.; Spearing, S. M.; Colombi Ciacchi, L. *J. Chem. Phys.* **2007**, *127*, 204704–204715.

(54) Lee, H.; Scherer, N. F.; Messersmith, P. B. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12999–13003.

(55) Ruoslahti, E.; Pierschbacher, M. D. *Science* **1987**, *238*, 491–497.

(56) Xiao, S.-J.; Textor, M.; Spencer, N. D. *Langmuir* **2998**, *14*, 5507–5516.

(57) Rammelt, S.; Illert, T.; Bierbaum, S.; Scharnweber, D.; Zwipp, H.; Schneiders, W. *Biomaterials* **2006**, *27*, 5561–5571.

(58) Song, D.-P.; Chen, M.-J.; Liang, Y.-C.; Bai, Q.-S.; Chen, J.-X.; Zheng, X.-F. *Acta Biomat.* **2010**, *6*, 684–694.

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

(60) Grossfield, A. *WHAM: The weighted histogram analysis method*; University of Rochester Medical Center: Rochester, NY; http://membrane.urmc.rochester.edu/content/wham. Accessed August 17, 2009.

(61) Darve, E. In *Free Energy Calculations*; Chipot, C., Pohorille, A., Eds.; Springer, Berlin, Germany, 2007; Chapter 4, pp 119−170.

(62) Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, W. F. *Chem. Phys. Chem.* **2007**, *8*, 162–169.

(63) Lee, I.; Marchant, R. *Surf. Sci.* **2001**, *491*, 433–443.

(64) Choi, Y.; Kim, E.; Lee, Y.; Han, M. H.; Kang, I.-C. *Proteomics* **2010**, *10*, 72–80.

(65) Csanyi, G.; Albaret, T.; Payne, M. C.; Vita, A. D. *Phys. Rev. Lett.* **2004**, *93*, 175503–175506.

CT1004388

JCTC Journal of Chemical Theory and Computation

First Principles Calculations of Atomic Nickel Redox Potentials and Dimerization Free Energies: A Study of Metal Nanoparticle Growth

Dian Jiao,[†] Kevin Leung,*[,‡] Susan B. Rempe,*[,†] and Tina M. Nenoff[‡]

Nanobiology Department, MS 0895, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States, and Surface and Interface Sciences Department, MS 1415, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States

Received August 3, 2010

Abstract: The redox potentials and dimerization free energies of transient transition metal cations in water shed light on the reactivity of species with unusual charge states and are particularly pertinent to understanding the mechanism and feasibility of radiolysis-assisted metal nanoparticle growth from salt solutions. A combination of quasi-chemical theory and *ab initio* molecular dynamics thermodynamic integration methods are applied to calculate these properties for nickel. The reduction potential for Ni^{2+} (aq) is predicted to be between −1.05 to −1.28 V, which is substantially lower than previous estimates. This suggests that Ni^{2+} reduction may possibly occur in the presence of organic radical anion electron scavengers and hydrogen atoms, not just hydrated electrons. In contrast, Ni^{+} is found to be stable against disproportionation. The formation of dimers Ni_2 and Ni_2^{+} from Ni and Ni^{+} are predicted to be favorable in water.

I. Introduction

Accurate modeling of reduction−oxidation (redox) reactions are pertinent to a wide range of electrochemical applications including batteries,[1] metal extraction,[2] catalysis,[3] and biology.[4,5] In particular, short-lived transition ionic metal species with unusual charge states are important intermediates in many multistep, multielectron processes. An intriguing application is the radiolysis-assisted synthesis of metal nanoparticles and alloys in aqueous solutions.[6−11] Secondary electrons from γ radiation or other sources can directly or indirectly (through electron-scavenging organic radical anions) reduce metal ions in salt solutions to their low oxidation states. Metal clusters are then formed via a series of reduction, clustering, and disproportionation reactions:[12−15]

$$\mathrm{Ni}_n^{q+} + \mathrm{e}^- \rightleftharpoons \mathrm{Ni}_n^{(q-1)+} \tag{1}$$

$$\mathrm{Ni}_n^{q+} + \mathrm{Ni}^{q'+} \rightleftharpoons \mathrm{Ni}_{n+1}^{(q+q')+} \tag{2}$$

$$\mathrm{Ni}^{q+} + \mathrm{Ni}^{q'+} \rightleftharpoons \mathrm{Ni}^{q''+} + \mathrm{Ni}^{(q''-q-q')+} \tag{3}$$

The choice of the Ni example allows us to support recent γ-irradiation experiments.[7,8] Not indicated in eqs 1−3 is the possibility of mixed alloy formation involving more than one metal element. In fact, the synthesis of unique alloys that are not thermodynamically stable can be accomplished via this route.[7,8]

The properties of transient metal species in water are difficult to measure. In this work, we use modeling techniques to investigate the initial stages of radiolysis-assisted Ni cluster formation demonstrated in experiments. The free energy change of clustering (eq 2) governs nanoparticle growth. The clustering of transition metal atoms is important to catalysis and has been extensively studied,[16−19] but the nucleation process has seldom been modeled in aqueous media. As a first step, we compute the dimerization free energies of Ni in water:

$$\begin{aligned} \mathrm{Ni}^{(0)} + \mathrm{Ni}^{(0)} &\rightleftharpoons \mathrm{Ni}_2^{(0)} \\ \mathrm{Ni}^{(0)} + \mathrm{Ni}^{+} &\rightleftharpoons \mathrm{Ni}_2^{+} \end{aligned} \tag{4}$$

* Corresponding author e-mails: slrempe@sandia.gov (S.B.R.), kleung@sandia.gov (K.L.).

[†] Nanobiology Department.

[‡] Surface and Interface Sciences Department.

10.1021/ct100431m
Published on Web 12/30/2010

Such clustering reactions have been measured for Ag,[20,21] but to our knowledge not the transition metal element Ni. Detailed studies of gas-phase, unhydrated Ni dimers have revealed much complexity.[22−29] Our work builds on these past studies but focuses on the interaction between the neutral and cationic monomers and dimers with liquid water. As will be shown, the formation of $Ni_2^{(0)}$ and Ni_2^{+} dimers are found to be favorable, and they represent the first steps of the complex mechanism in the formation of Ni nanoclusters from aqueous salt solutions.

We also consider another significant property that can benefit from theoretical input, namely, the redox potentials (Φ_{redox}) associated with transient Ni^{q+} and Ni_2^{q+} species. Φ_{redox} plays a crucial role in determining what cluster size, stoichiometry, and net charge can exist under experimental conditions. As discussed in pioneering work on the radiolysis synthesis of Au, Ag, and bimetallic Au/Ag nanoclusters,[30,31] the detailed mechanism of cluster formation depends on the redox potentials of the metal salts in solution, in our case Ni_n^{q+} (eq 1) and analogous bimetallic species relative to the reducing species (excess electrons and organic radical anions) that exist in the γ-irradiated solution. The excess electron chemical potential in liquid water is generally listed at −2.7 to −2.9 V relative to the standard hydrogen electrode (SHE),[32,33] while electron-scavenging hydroxymethyl radical anions present in radiolysis experiments are at a more modest −1.18 V.[7] For Ni_n^{q+} to be reduced in solution, it must exhibit a more positive Φ_{redox} than the relevant redox potential of the electron-donating species.

Φ_{redox} can be rigorously separated into two contributions: the change in standard state ion hydration free energy (ΔG_{hyd}) and the ionization potential (IP). Φ_{redox} is referenced to the SHE by subtracting 4.44 V. In the literature, the reported Φ_{redox} of transient first row transition metal ions in water often contain theoretical components. Thus, using experimental IP and Ni^{+} and Ni^{2+} ΔG_{hyd} values estimated via Pauling radius interpolation and the Born hydration free energy formula,[34] respectively, Baxendale and co-workers[13,14] have reported a −2.7 V Φ_{redox} for $Ni^{2+} + e^{-} \rightleftharpoons Ni^{+}$. This Φ_{redox} is very close to that of the electron injected into water,[32,33] suggesting that hydrated electrons are marginally sufficient to reduce Ni^{2+}. At the −1.18 V associated with organic radical anions found under radiolysis conditions,[7] Ni^{2+} should be inert. Baxendale et al. have, however, ignored the ligand-field splitting induced energetic stabilization arising from the first hydration shell water molecules, which can amount to a significant fraction of an electronvolt for first row transition metal ions.[35] As will be shown, our predictions lead to a substantial revision of this earlier estimate.[13]

Two modern computational strategies have been applied to calculate ΔG_{hyd}. In the more widely used approach, the Density Functional Theory (DFT) electronic structure method is used to optimize the geometry of gas-phase clusters of transition metal ions containing first hydration shell water molecules.[36−42] DFT explicitly takes into account ligand field splittings. The outer shell water molecules can be represented explicitly or by an implicit solvent model via a dielectric continuum approximation.[40,42] A successful implementation of this solvation method[35] is the "quasi-chemical theory" (QCT).[43−46] This method makes use of the most probable distribution of hydration numbers (i.e., the number of water molecules, N_w, residing in the ionic hydration shell). The predicted equilibrium hydration number can be compared with X-ray and neutron scattering data and is complementary to nuclear magnetic resonance relaxation time information.[47]

A second approach applies *ab initio* molecular dynamics (AIMD) simulations, where metal ions and all water molecules, including outer-shell ones, are treated explicitly using DFT at each finite temperature molecular dynamics time step.[48−51] Thermodynamic integration[52] (TI) using AIMD simulations have yielded hydration free energies for simple ions in good agreement with experiments,[53,54] suggesting that reliable Φ_{redox} can also be predicted with the AIMD method provided an accurate IP can be obtained theoretically or experimentally.

The two theoretical methods complement each other. The QCT approach directly computes the hydration free energy of any species, divided into contributions that provide insights into the effects from local and distant solvent. Past work attests to the success of this approach in calculating hydration free energies of small molecules.[35,43,44,55] The approach has also been used to investigate mechanisms of selectivity in biological ion binding sites.[56,57] The more costly AIMD method can generate new insights into the bulk solvation structures of low-valence transition metals, which are of basic scientific interest due to their unusual electronic properties.[50,58−61] For transition metal elements, AIMD readily yields differences in ΔG_{hyd} between different ionic charge states, but not the ΔG_{hyd} themselves. A comparison between AIMD and QCT ΔG_{hyd} changes as reduction reactions occur allows critical examination of the different approximations used in both approaches and helps elucidate the discrepancies between theoretical and experimental Φ_{redox} reported in the literature.[37−39,41]

Treatment of the transition metal ion d electrons is efficiently improved using DFT+U (Hubbard-like) augmentations[62,63] within AIMD simulations, which has been applied to molecular systems.[64] QCT has the advantage of being much more computationally efficient and permits the use of hybrid DFT functionals, which are generally more accurate than non-Hubbard augmented, nonhybrid DFT functionals for the properties of main group elements and many transition metal complexes. The coupled-cluster (CCSD(T)) level of theory, more reliable than hybrid functionals, can also be applied to calculate gas-phase binding energies and calibrate DFT results.

Here, we have applied first principles methods and the theoretical frameworks described above to calculate the redox potentials and dimerization free energies of monomeric and dimeric nickel species in liquid water. In the following parts of this paper, section II details the AIMD and QCT methods used. Section III describes the computational results, and section IV concludes the paper with brief discussions.

II. Method

II.A. Hydration Calculation by QCT. The quasi-chemical theory enables calculation of the hydration free energy in terms of individual contributions from inner-shell and outer-shell solvent domains. The inner-shell domain typically consists of the ion and water molecules that form the first hydration shell. The binding free energy for the formation of inner-shell complexes can be computed in the absence of outer-shell solvent as

$$\Delta G^{(0)} = G^{(0)}_{\mathrm{Ni}^{q+}(\mathrm{H_2O})_n} - G^{(0)}_{\mathrm{Ni}^{q+}} - nG^{(0)}_{\mathrm{H_2O}} \quad (5)$$

The equilibrated structures of nickel−water clusters, which represent the inner hydration shell regions, are optimized using DFT methods starting from AIMD simulation configurations. QCT DFT calculations are performed using the Gaussian suite of programs.[65] Frequencies and zero-point energies are determined using CCSD(T)[66] and the Becke-3-parameter-Lee−Yang−Parr (B3LYP) functional.[67] Some results for the Perdew−Burke−Ernzerhof (PBE) functional[68] are also found in the Supporting Information document (SI). The 6-311+G(d,p) basis set is applied throughout. Standard, finite temperature and zero point energy contributions are added to ΔG^0. The optimal spin state is chosen as the most stable state.

Free energy contributions from the outer region consist of the molecular packing contribution (cavity) and the interactions between the inner-shell cluster and outer-shell solvent molecules. Since AIMD calculations do not take into account the small cavity contribution, this term is left out from the QCT calculation for the sake of better comparison with AIMD. The solvation effects from the outer region are obtained by treating the external solvent as a dielectric continuum. The electrostatic potential is evaluated by solving the Poisson−Boltzmann equation with the APBS package.[72] The numerical technique used to solve the equation is a combination of the standard finite difference focusing method and the parallel adaptive finite element algorithm.[73] For the APBS calculation, partial charges on an inner-shell complex are acquired from the ChelpG method with the 6-311+G(d,p) basis set, while the radii for oxygen and hydrogen atoms are taken from the literature.[69] Radii of nickel ions used to define the division between inner and outer solvent domains are determined by the first minima in ion−oxygen pairwise correlation functions, $g(r)$, from AIMD simulations. Where experimental data is available (e.g., Ni^{2+}), the simulated minima match results from X-ray experiments.[70,71] The fine mesh domain length is set to 10 Å and the coarse mesh domain length, 30 Å. The dielectric constants for the inner and outer shells are set to 1.0 and 78.5, respectively.

II.B. AIMD Simulations. Spin-polarized AIMD simulations apply the Vienna *ab initio* simulation program (VASP),[74] projected-augmented wave (PAW) pseudopotentials[75,76] (PP) with only valence electrons for H and O atoms, and a Ni PP that includes pseudovalent 3p electrons, the PBE functional,[68] Γ-point Brillouin zone sampling, and a 400 eV plane-wave energy cutoff.

Semilocal functionals such as PBE are generally inadequate for treating first row transition metal complexes,[77] even while they can be successful with dimers in the gas phase.[78] As will be shown, the B3LYP functional is more accurate for depicting interactions of Ni species with water. Since the hybrid functional B3LYP is too costly to use in AIMD settings, we apply the DFT+U approach[62] to Ni 3d orbitals only. The U value is fitted to reflect the zero temperature B3LYP binding energy in the $Ni^{2+}(H_2O)_6$ cluster. AIMD trajectories for $q \leq 1$ are generated using the PBE functional, whereas Ni^{q+} (aq) trajectories for $q \geq 1$, taken from ref 53, are generated using the DFT+U functional with $U = 4$ eV. Fitting VASP to Gaussian results is possible because, at $T = 0$ K, PBE geometry optimization calculations that apply VASP PAW pseudopotentials and the plane-wave basis and those that apply the Gaussian suite of codes and the 6-311+G(d,p) basis[65] yield $Ni^{2+}(H_2O)_6$ binding energies that agree to within a few tenths of an electronvolt. While using two types of functionals to generate AIMD trajectories is admittedly awkward, the hydration free energies should not be strongly affected. As discussed in ref 53, the hydration *structures* of most Ni species are sufficiently similar for the PBE and DFT+U methods, so that using either to generate trajectories should yield very similar ΔG_{hyd} values. An estimate of the small error introduced in using AIMD/PBE trajectories will be discussed in the Results section and in the SI. (In contrast, DFT+U and PBE predict ligand field splittings and reaction *energies* that differ by fractions of an electronvolt.) PBE is selected to generate trajectories in this work because, for the Ni_2^{2+} (aq) species only, it gives stable structures, while B3LYP does not (see below).

A Nose thermostat fixes the temperature at $T = 400$ K, which is needed for the PBE functional to describe the room temperature experimental liquid water structure.[79] The deuterium mass is adopted for all protons to allow a larger time step while the H mass is assumed whenever water density is reported. Along with Born−Oppenheimer dynamics time steps of 0.25 fs and a 10^{-6} eV energy convergence criterion, these settings limit the temperature drifts to 1 K/ps. The trajectory length is 30 ps for each of the TI windows. Initial configurations are pre-equilibrated using the extended simple point charge (SPC/E) water model[80] and a Ni^{q+} force field consisting of a $+q$ point-charge scaled to the net charge of the corresponding AIMD simulation cell plus a Lennard-Jones functional form. Such crude force fields do not yield the well-structured first hydration shells of transition metal ions but are useful for dielectric relaxation of outer-shell water molecules that accompany changes in ionic charges. After switching from force fields to AIMD simulations, we find that the distinctive Ni^{q+} first hydration shell generally becomes equilibrated and yields the expected structures within 2 ps, except for $Ni_2^{(0)}$, which takes 7 ps to reach the equilibrium hydration number (N_w). The short equilibration time suggests that AIMD predicted free energy changes should not depend on initial conditions.

AIMD simulations for a single Ni^{q+} ion, $0 \leq q \leq 2$, are performed using $9.885 \times 9.885 \times 9.885$ Å^3 simulation cells that contain a Ni atom/ion and 32 H_2O molecules. In the case of the dimer, the $12.885 \times 9.885 \times 9.885$ Å^3 cells contain a Ni dimer and 40 water molecules, with the x and

y coordinates of both Ni atoms held fixed and identical along the long axis of the simulation cell.

II.C. Hydration Free Energy Changes via Thermodynamic Integration. To calculate differences in hydration free energy, $\Delta\Delta G_{hyd}$, between species in different charge states (q_i and q_f) using AIMD simulations, thermodynamic integration (TI)[52,81] was performed via

$$\Delta\Delta G_{Hyd} = \int_{q_i}^{q_f} \langle dH(\lambda)/d\lambda \rangle_\lambda \, d\lambda + (q_f - q_i)\Delta\phi \qquad (6)$$

Here, $H(q)$ is the total potential energy of the simulation cell at a net charge $\lambda = q$ in the simulation cell computed using a modified version of VASP,[82] minus the energy of the isolated Ni^{q+} with the same spin state. In VASP simulations, q is set by imposing a fixed number of electrons (which can be a noninteger) in the simulation cell. This is appropriate for modeling hydrated $Ni^{(0)}$, Ni^{+}, Ni^{2+}, and the intermediate charges bracketed by these species because the fractional electron is found to reside on the Ni 3d orbitals via maximally localized Wannier functional analysis.[85] The one exception to this rule will be discussed in section III.D. The net spin is tuned to a value linearly interpolated between the stable spin state of the end points of the TI simulation in the aqueous phase, *not* necessarily those of the stable spin states of the gas-phase Ni species. This procedure captures the net hydration free energy, including the ligand field splitting, to be discussed in more detail in section III.C. It excludes the gas-phase ionization potential contribution, which tends not to be predicted accurately from widely used DFT functionals. The IP value is taken from experiments whenever possible. If the stable spin states of the bare and hydrated Ni species differ, as is the case with $Ni^{(0)}$, a spin-flip energy is added post processing to the change in hydration free energy.

Also included in $H(q)$ are monopole- and image charge-induced electrostatic corrections due to the periodic boundary conditions,[83] $\Delta E_{Ewald} = q^2/(2\alpha\varepsilon_o L)$, where α is the Madelung constant, L is the cubic cell size dimension, and ε_o is the pertinent dielectric constant. ε_o is set to unity for isolated ions and to infinity for water. Another, much smaller finite size correction, the change in $\Delta E_{quad} = 2\pi q^2 R^2/3L^3$, where R is the ion radius, is added to $\Delta\Delta G_{hyd}$ post processing.[84] With our simulation cell size and an estimated $R = 2$ Å for Ni^{q+}, which reflects the approximate position of the peaks in $g(r)$ (Figures 1 and 2), ΔE_{quad} amounts to 0.125 and 0.500 eV for $q = 1$ and 2, respectively.

Operationally, at every 0.1 ps interval, we use finite difference to sample $dH(q)/dq \approx [H(q_+) - H(q_-)]/(|e|/20)$, $q_\pm = q \pm |e|/40$, at the fixed atomic configurations in the snapshot. When q is an integer, $q_\pm$ values are shifted so that they do not exceed the boundary values of the electron addition half-cell reaction. A six-point trapezoidal rule integrates over the resulting $\langle dH(q)/dq \rangle_q$. Equation 6 is rigorous even for models with electronic polarizability.

Surface potentials ($\Delta\phi$) at fluid−fluid and fluid−solid interfaces contribute to solvation free energies. In particular, the air−water surface potential is an integral part of the ion hydration free energy ΔG_{hyd} via the second term, $(q_f - q_i)\Delta\phi$, in eq 6.[86] $\Delta\phi$ can be decomposed into dipolar ($\Delta\phi_d$) and quadrupolar ($\Delta\phi_q$) contributions.[86] $\Delta\phi_q$ is a bulk liquid water quantity independent of the nature of interfaces, and the theoretical $\Delta\phi_q$ value has been estimated using the PBE quadrupole component at 1.0 g/cm^3 water density.[53,87] For $\Delta\phi_d$, we note that this work seeks to mimic electrochemical measurements of electron transfer between a metal electrode and Ni^{q+} species in water. To our knowledge, the pertinent water-electrode $\Delta\phi_d$ has not been computed using atomic simulations. The infinite dielectric constant of metal, which leads to "image charges" inside the metal,[88] should partly compensate for the potential drop due to water surface dipoles. It should be a reasonable assumption that the water−metal electrode ϕ_d is smaller in magnitude than the air−water ϕ_d. Thus, we have omitted $\Delta\phi_d$ in our $\Delta\Delta G_{hyd}$ calculation. This argument is not completely rigorous because both $\Delta\phi_d$ and $\Delta\phi_q$ depend on the choice of molecular center; only $\Delta\phi$ is independent of such choices. Nevertheless, from

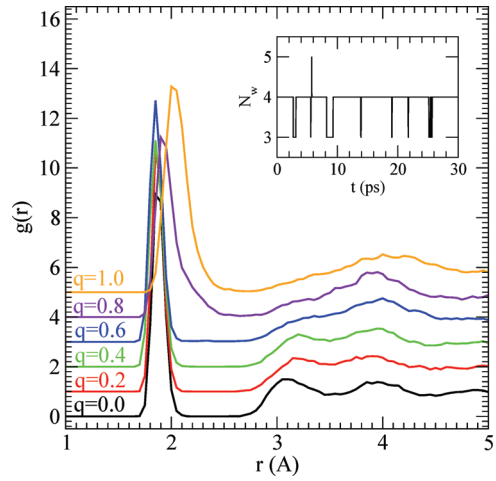


Figure 1. AIMD pair correlation function $g(r)$ between O_w and Ni^{q+} as q varies. Black, red, green, blue, violet, and orange lines denote $q = 0.0$, 0.2, 0.4, 0.6, 0.8, and 1.0, respectively, offset by one density unit along the y axis. $N_w = 2.00$, 2.00, 2.03, 2.06, 2.86, and 3.93 for these Ni^{q+} species, respectively. The inset depicts instantaneous N_w for $q = 1.0$.

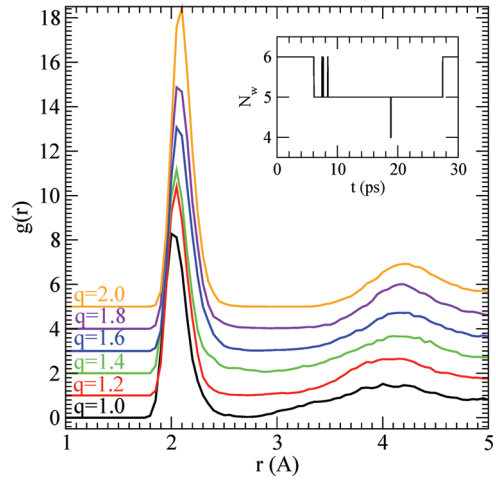


Figure 2. AIMD pair correlation function $g(r)$ between O_w and Ni^{q+} as q varies. Black, red, green, blue, violet, and orange lines denote $q = 1.0$, 1.2, 1.4, 1.6, 1.8, and 2.0, respectively. $N_w = 3.81$, 4.34, 4.80, 5.17, 5.48, and 6.00 for these q values, respectively. The inset depicts the instantaneous N_w for $q = 1.8$ (black).

the image-charge argument, the ambiguity in redox potential introduced from omitting ϕ_d should be less than the 0.21 V associated with vapor-water $\Delta\phi_d$ for the SPC/E water model. Both Baxendale et al.'s estimates[13,14] and the QCT approach (see below) exclude the surface potential.

III. Results

III.A. Choice of DFT Functional and Method. First, we describe the benchmarking procedure that informs our choice of DFT functional. The gas-phase $Ni^{(0)}(H_2O)_2$ cluster is sufficiently small so that the more reliable quantum chemistry CCSD(T) method can be used to perform single point binding energy calculations to calibrate DFT results. PBE, B3LYP, and CCSD(T) predict an ΔE_{bind} of 1.88, 1.06, and 1.28 eV, respectively, for this complex. This test suggests that B3LYP is more reliable than PBE for $Ni^{(0)}$ hydration.

PBE and B3LYP predict 16.31 and 15.71 eV binding energies for the gas-phase $Ni^{2+}(H_2O)_6$ cluster. The 0.60 eV discrepancy is mostly due to differences in energies from ligand field splitting, which was estimated in ref 35 as follows. The QCT/B3LYP ΔG_{hyd} values are computed for Ca^{2+} and Zn^{2+}, which contain either an empty or a full 3d shell and therefore do not exhibit ligand field splitting. Interpolating between these extremes and examining the deviation of individual transition metal ions as a function of 3d orbital occupation leads to a 1.3 eV ligand-field stabilization for Ni^{2+} in water. We have reproduced a similar result by considering only the B3LYP gas-phase $M^{2+}(H_2O)_6$ energy at $T = 0$ K, excluding zero-point corrections, outer-shell water contributions, and thermal effects. We conclude that gas-phase $Ni^{2+}(H_2O)_6$ is stabilized by 1.40 eV relative to the value interpolated between Ca^{2+} and Zn^{2+}. Using this gas-phase route, the PBE functional yields a 1.80 eV ligand-field stabilization energy for Ni^{2+}. The B3LYP ligand-field splitting is in better agreement with the 1.26 eV spectroscopic data[35,89] than the PBE one. Thus, B3LYP should be considered more accurate for Ni^{2+} hydration.

PBE and B3LYP binding energies for gas-phase $Ni^{+}(H_2O)_4$ also differ by 0.48 eV. An effort to interpolate ligand-field stabilization for the $Ni^{+}(H_2O)_4$ complex fails. While $K^{+}(H_2O)_4$, devoid of 3d electrons, is stable in the gas phase, $Cu^{+}(H_2O)_4$, the hydrated species, which could have fully occupied 3d orbitals and no 4s electrons, collapses. This complex turns into a linear, two-coordinated Cu with the two other water molecules relegated to the outer shell and linked to the two inner shell H_2O's through hydrogen bonds. The linear structure is consistent with previous AIMD simulations of hydrated Cu^{+}.[59]

Despite lacking an estimate of ligand-field stabilization energy for $Ni^{+}(H_2O)_4$ to validate PBE and B3LYP binding energies, we can test binding between water and neutral and divalent nickel atoms. PBE is shown to substantially overestimate the binding between water and both $Ni^{(0)}$ and Ni^{2+}. Thus, in the remainder of this manuscript, we focus on two complementary methods, QCT/B3LYP and AIMD/DFT+U with the Hubbard U value fitted to the B3LYP $Ni^{2+}(H_2O)_6$ binding energy.

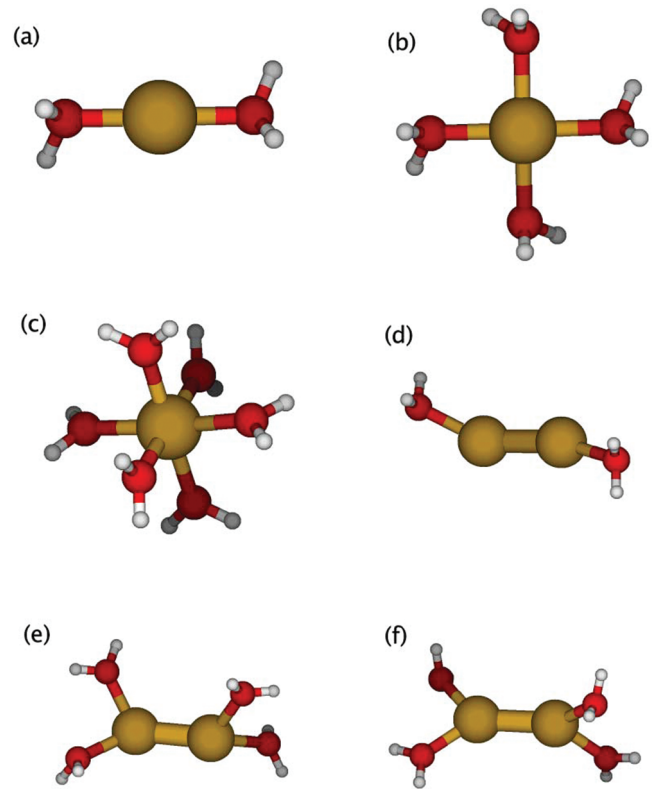


Figure 3. Optimized gas-phase clusters using the Gaussian code and B3LYP/6-311+G(d,p) level of theory. (a) $Ni^{(0)}(H_2O)_2$ ($s = 0$); (b) $Ni^{+}(H_2O)_4$ ($s = 1/2$); (c) $Ni^{2+}(H_2O)_6$ ($s = 1$); (d) $Ni_2{}^{(0)}(H_2O)_2$ ($s = 1$); (e) $Ni_2{}^{+}(H_2O)_4$ ($s = 3/2$); (f) $Ni_2{}^{+}(H_2O)_4$ ($s = 1/2$). Blue, red, and white spheres represent Ni, O, and H atoms, respectively.

III.B. Hydration Structures. Figures 1 and 2 depict the pair correlation functions, $g(r)$, between Ni^{q+} and the water oxygen (O_w) site. Only integral values of q are physical, but the fractional q results, needed for TI calculations, reveal interesting trends in the hydration structure. Also listed in the figure captions are the hydration numbers, N_w, defined as the spatial integral up to the first minimum ($r = 2.6$ Å) in $g(r)$.

While the bare Ni atom is a spin triplet ($s = 1$), both B3LYP and DFT+U predict that $Ni^{(0)}(H_2O)_2$ is most stable in the singlet state ($s = 0$) in the gas phase. (See the SI for more spin state information.) The $g(r)$ between singlet $Ni^{(0)}$ and the oxygen site of water molecules obtained in AIMD simulations exhibits a sharp first peak at $R_{Ni-O} = 1.8$ Å that integrates to two H_2O molecules. The instantaneous hydration configuration in liquid water is linear, similar to the gas-phase optimized structure shown in Figure 3a, which in turn resembles that of $Cu^{+}(H_2O)_2$.[59] The sharp first peak reflects strong covalent bonds between water and $Ni^{(0)}$. As q increases from 0.0 to 1.0, the first peak in $g(r)$ broadens and shifts to a larger distance of $r = 2.0$ Å. This is presumably because more water molecules enter the hydration shell at larger q. The increased repulsion between the oxygen sites of the first shell hydration H_2O molecules should weaken the interaction between each individual H_2O and the Ni^{q+}. At $q = 1.0$, the number of H_2O molecules in the first hydration shell increases to 3.8 and a stable square planar hydration shell is formed, just like the gas-phase $Ni^{+}(H_2O)_4$

complex (Figure 3b). The secondary structure in $g(r)$ at $r =$ 3.1 Å, clearly discernible at $q = 0$, also becomes smeared out at larger q and disappears beyond $q = 0.4$.

Ni^+ is in the doublet ($s = 1/2$) spin state with this ligand field. The inset to Figure 1 depicts the temporal fluctuations in N_w at $q = 1.0$, showing that the water molecules enter and leave the first hydration shell on a sub-picosecond time scale despite the stable hydration structure. Two of the H_2O molecules initially residing in the hydration shell have been replaced by H_2O molecules from the outlying regions by the end of this 30 ps trajectory.

As q further increases from 1 to 2 (see Figure 2), N_w smoothly rises from 4 to 6. The first peak position in $g(r)$ remains at $r \approx 2.1$ Å but sharpens as q approaches 2, at which point the well-known Ni^{2+} octahedral first hydration shell[35] (Figure 3c) is observed in AIMD simulations. Fluctuations of instantaneous N_w also occur on sub-picosecond time scales for $q < 2$ except at $q = 1.8$, where transition between 5- and 6-fold coordination occurs more slowly (inset to Figure 2).

Turning our attention to the dimers, $Ni_2^{(0)}$ is predicted to be a spin triplet with all DFT methods we have considered. In AIMD simulations of this species in water, the Ni atoms collectively exhibit $N_w = 3$; one Ni is typically instantaneously coordinated to two H_2O molecules and the other to one H_2O. In the gas phase, two-coordinated $Ni_2^{(0)}(H_2O)_2$ forms a stable linear cluster (Figure 3d). However, three-coordinated $Ni_2^{(0)}(H_2O)_3$ is not stable; one of the H_2O molecules migrates to the outer shell, forming a hydrogen bond with one of the two remaining H_2O's directly coordinated to Ni. Thus, for QCT/B3LYP calculations, we focus on the two-coordinated $Ni_2^{(0)}$.

For the Ni_2^+ cation dimer, the B3LYP functional predicts that the quartet ($s = 3/2$) state is more stable than the doublet ($s = 1/2$) with and without dielectric continuum treatment of outer-shell water. Two H_2O molecules are coordinated to each Ni of the $Ni_2^+(H_2O)_4$ complex. The two Ni atoms exhibit similar O_w-Ni-O_w angles (Figure 3e) and equal net integrated charge and spin densities according to Mulliken analysis.[90] Figure 4 depicts the $g(r)$ between Ni and O_w for quartet Ni_2^+.

For completeness, we also briefly discuss the doublet $Ni_2^+(H_2O)_4$ cluster, which is metastable in the gas phase. The two Ni atoms are in manifestly different chemical bonding environments (Figure 3f). Mulliken analysis reveals that one Ni has a large net charge and spin density, and appropriately the two coordinating water molecules are at 90° to each other as in a square-planar $Ni^+(H_2O)_4$ complex. The other Ni is charge neutral, has little or no spin density, and the two H_2O's are indeed at 180° from each other as in the linear $Ni^{(0)}(H_2O)_2$ complex. Such solvent-induced asymmetry in charge distribution has been examined in I_3^- and other systems.[91] This low-spin cluster will not be the subject of AIMD free energy calculations.

$Ni_2^{2+}(H_2O)_n$ gas-phase complexes are unstable within the B3LYP treatment; the two Ni^+'s become separated by a water molecule during geometry optimization. In contrast, in PBE calculations, the Ni−Ni bond does not spontaneously break in $Ni_2^{2+}(H_2O)_5$. Consistent with this gas-phase prediction, the Ni_2^{2+} complex is stable in AIMD/PBE simulations. Recall this is the reason we choose to run AIMD simulations with the PBE functional and then perform DFT+U calculations for the hydration free energy calculations based on PBE configurations. The two Ni's in Ni_2^{2+} yield a combined $N_w = 5.22$. Individually, each Ni is found to exhibit $N_w \approx 3$; thus on average, one H_2O is shared between the two Ni, simultaneously within the hydration shell of both. This sharing does not occur for $Ni_2^{(0)}$ and Ni_2^+ but is indeed observed in the gas-phase $Ni_2^{2+}(H_2O)_5$ optimized geometry (not shown). Mulliken analysis shows that both Ni ions in this cluster have equal integrated charge and spin densities. Since the covalently bonded Ni_2^{2+} dimer is only stable in water when using the PBE functional, it is likely only marginally stable, and its redox properties will not be the subject of this work.

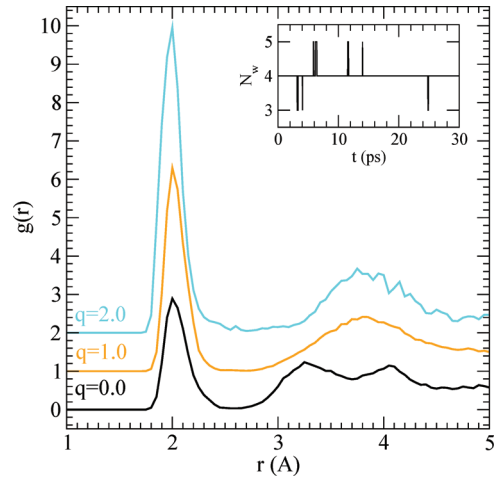


Figure 4. $g(r)$ between O_w and the Ni sites for Ni_2^{q+} as q varies. Black, orange, and cyan lines denote $q = 0.0$, 1.0, and 2.0, respectively. $N_w = 3.00$, 4.01, and 5.22 for the two Ni atoms combined. The inset depicts instantaneous N_w for $q = 1.0$.

The hydration number predictions from AIMD simulations for Ni^{q+} and Ni_2^{q+} species are used to determine the number of water molecules used in gas-phase cluster calculations on which the QCT method is based.

III.C. QCT Hydration Free Energies. The components of the QCT absolute hydration free energies for $Ni^{(0)}$, Ni^+, and Ni^{2+} are listed in Table 1. It is worth noting that the cavitation energy is not included. However, we estimate the packing contribution to the solvation based on the volume of the cavity. The work that is done to create a cavity by solvating the ion is 0.06, 0.15, and 0.13 eV for $Ni^{(0)}$, Ni^+, and Ni^{2+}, respectively, which is insignificant in comparison to the total hydration free energy. The absolute $\Delta G_{hyd} = -20.59$ eV calculated using the B3LYP functional for Ni^{2+} hydration is in reasonable agreement with experiments (−20.79 eV)[92] and the prediction in ref 35 using similar methods (−20.49 eV). QCT/B3LYP predicts $\Delta G_{hyd} = -5.81$ for Ni^+. Obviously, with the increment of the net charge to higher values, the hydration free energy increases to more negative values. By comparison, Dixon and Baxendale[13] adopted ΔG_{hyd} values of −4.90 and −21.39 eV for Ni^+ and Ni^{2+}, which are significantly different from the QCT/B3LYP results.

Table 1. Hydration Free Energies of Nickel Species Calculated Using QCT/B3LYP at $T = 300$ K, except That the Asterisk Indicates a CCSD(T) Calculation[a]

reactions	$\Delta G^{(0)}$	ΔG	$\Delta\mu$	ΔG_{hyd}
$Ni^{(0)} + 2H_2O \rightleftharpoons Ni^{(0)}(H_2O)_2$	−0.217	−0.588	0.163	−0.425
$*Ni^{(0)} + 2H_2O \rightleftharpoons Ni^{(0)}(H_2O)_2$	−0.403	−0.774	0.210	−0.564
$Ni^{+} + 4H_2O \rightleftharpoons Ni^{+}(H_2O)_4$	−3.757	−4.497	−1.310	−5.807
$Ni^{2+} + 6H_2O \rightleftharpoons Ni^{2+}(H_2O)_6$	−12.819	−13.930	−6.658	−20.588
$Ni_2^{(0)} + 2H_2O \rightleftharpoons Ni_2^{(0)}(H_2O)_2$	−0.677	−1.047	0.129	−0.919
$Ni_2^{+} + 4H_2O \rightleftharpoons Ni_2^{+}(H_2O)_4$ ($s = 3/2$)	−3.463	−4.203	−1.139	−5.342
$Ni_2^{+} + 4H_2O \rightleftharpoons Ni_2^{+}(H_2O)_4$ ($s = 1/2$)	−3.006	−3.746	−1.132	−4.878
H_2O				−0.361

[a] $\Delta G^{(0)}$ is the free energy change for formation of inner-shell clusters in the absence of the surrounding medium. $\Delta G = \Delta G^{(0)} - nRT$ ln(1354) is the free energy change accounting for the actual density of water. $\Delta\mu$ is the electrostatic interaction between clusters in the inner shell and the implicit solvent in the outer shell. Combining these contributions yields ΔG_{hyd}, the standard state hydration free energy of the Ni_n^{q+} species. Hydration free energy of the water molecule estimated using partial charges from CCSD(T) is −0.390 eV. All values are given in electronvolts.

Table 2. Hydration Free Energy Differences between Nickel Species in Different Charge States ($\Delta\Delta G_{hyd}$, in eV) and Redox Potentials (Φ_{redox}, volt) Estimated by QCT and AIMD[a]

	method	functional	$\Delta\Delta G_{hyd}$	Φ_{redox}
$Ni^{+} \rightleftharpoons Ni^{(0)}$	QCT	B3LYP	5.38	−2.18
$Ni^{+} \rightleftharpoons Ni^{(0)}$	AIMD	DFT+U	5.37	−2.17
$Ni^{2+} \rightleftharpoons Ni^{+}$	QCT	B3LYP	14.78	−1.05
$Ni^{2+} \rightleftharpoons Ni^{+}$	AIMD	DFT+U	15.01	−1.28
$Ni_2^{+} \rightleftharpoons Ni_2^{(0)}$	QCT	B3LYP	4.42	−1.44

[a] $\Phi_{redox} = (-\Delta\Delta G_{hyd}/|e| + IP - 4.44)$ volt. The IP for Ni, Ni^{+}, and Ni_2 is 7.640, 18.170, and 7.420 eV, respectively. The AIMD $\Delta\Delta G_{hyd}$'s for Ni^{+} and Ni^{2+} exhibit standard deviations of 0.02 and 0.04 eV (0.5 and 1.0 kcal/mol), respectively.

The QCT/B3LYP method predicts $\Delta G_{hyd} = -0.43$ eV for $Ni^{(0)}$. The QCT/CCSD(T) method gives a similar result for $Ni^{(0)}$, $\Delta G_{hyd} = -0.56$ eV. Of all of the charge states, only the solvation of Ni^{2+} has been experimentally documented; other species have only a transient existence. Although carrying zero charge, the neutral Ni monomer and dimer both have significantly favorable solvation in water. Thus, unlike the noble Ag atom dispersed in water,[53] reactive transition metal atoms like Ni cannot be described as inert Lennard-Jones spheres.

III.D. Hydration Free Fnergy Changes upon Reduction. Table 2 depicts the difference in hydration free energies ($\Delta\Delta G_{hyd}$) between species differing by one electron. Note that Dixon and Baxendale's $\Delta\Delta G_{hyd}$ estimate for $Ni^{+} \rightarrow Ni^{2+}$ is 16.5 eV, which is at least 1.5 eV (~34 kcal/mol) higher than our DFT-based estimates. The standard deviations estimated for these AIMD $\Delta\Delta G_{hyd}$ are 0.02 (0.5 kcal/mol) and 0.04 eV (1.0 kcal/mol), respectively. They are calculated by splitting the trajectory in each window into four equal parts, computing the $\Delta\Delta G_{hyd}$ of each of the four segments, calculating the standard deviation of each, and dividing by $\sqrt{4}$.

Unlike the Li^{+} ion, which can be scaled to zero in its entirety,[53] transition metal ions carry a substantial amount of d-shell electrons, and AIMD cannot readily be used to yield ΔG_{hyd}.[53] The $\Delta\Delta G_{hyd}$'s are, however, obtained in a straightforward manner. AIMD/DFT+U can be compared to QCT/B3LYP results because the U parameter is fitted to the B3LYP $Ni^{2+}(H_2O)_6$ binding energy.

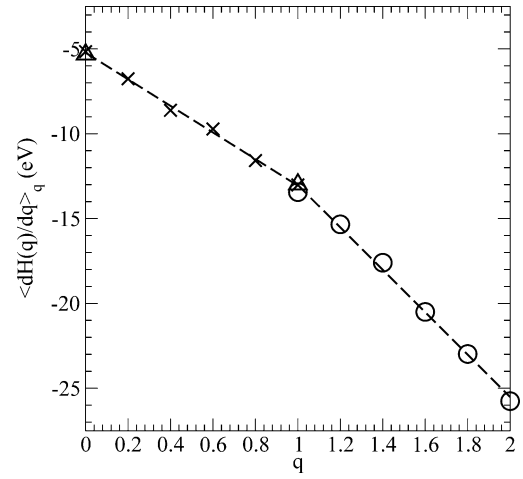


Figure 5. $\langle dH(q)/dq\rangle_q$ as a function of q in Ni^{q+} computed using the DFT+U method. Crosses are for $q \leq 1$, based on configurations generated using AIMD/PBE trajectories, while circles for $q \geq 1$ use snapshots taken from AIMD/DFT+U trajectories of ref 53. The triangles are test cases that also apply AIMD/DFT+U to generate trajectories in addition to evaluating $\langle dH(q)/dq\rangle_q$. The results are similar where AIMD/PBE trajectories (crosses) are applied. See text for further discussions.

Reasonable agreement is obtained between QCT and AIMD $\Delta\Delta G_{hyd}$'s, although different approximations are applied. AIMD conducts molecular dynamics sampling of nuclear motion using Newtonian dynamics and does not contain zero-point energies (ZPEs). The QCT approach approximates thermal effects with the harmonic approximation but includes quantum nuclear motion. We have, however, shown that ZPEs contribute minimally, on the order of 0.04 eV (1 kcal/mol), to $\Delta\Delta G_{hyd}$ (SI). The harmonic approximation is expected to be excellent for rigid hydration structures as in the case with the highly charged, octahedral $Ni^{2+}(H_2O)_6$. QCT treats the outer shell water molecules as a dielectric continuum. This does not appear to introduce significant discrepancy.

The AIMD TI calculations are worth examining in more detail. Figure 5 depicts $\langle dH(q)/dq\rangle_q$, which are the integrands in TI for AIMD ΔG_{hyd} as q varies. Despite the large ligand-field splitting and the jumps between the well-defined linear, square-planar, and octahedral hydation structures with the Ni coordinated to two, four, and six H_2O molecules, $\langle dH(q)/dq\rangle_q$ remains reasonably linear, especially for $0.2 \leq q \leq 0.8$ and $1 \leq q \leq 2$. This implies that the six-point trapezoidal discretization is adequate. In fact, a two-point integration scheme already yields a ΔG_{hyd} to within 1 kcal/mol. This behavior seems to be in contrast to an AIMD simulation of the energy gap for $Cu^{+} \rightarrow Cu^{2+}$, which has been shown to be nonlinear.[50] The relative lack of curvature in Figure 5 may arise from the exclusion of the bare Ni^{q+} energy. Consequently, $\langle dH(q)/dq\rangle_q$ only reflects hydration effects. The bare Ni^{q+} energy may be the main cause of nonlinearity in ref 50 as q varies because of self-interaction errors in approximate DFT functionals.[93] We also stress that $\langle dH(q)/$

$dq\rangle$, the rigorous integrand in TI used here, cannot be treated as vertical reorganization energy. The largest numerical uncertainties occur at the crossover regions where the hydration number changes significantly. For example, the slow fluctuations in N_w for $Ni^{1.8+}$ (inset to Figure 2) yield the most significant numerical noise in the ΔG_{hyd} integration.

The integrands $\langle dH(q)/dq\rangle_q$ for $q < 1$ and $q > 1$ do not necessarily match at $q = 1$. Their slopes with respect to q also differ. Even in the gas phase, adding or removing a small fractional electron to Ni^+ should yield different results because the electron affinity and ionization potential of Ni^+ differ. As alluded to in section II.C, Wannier function analyses of AIMD snapshots show that all but one value of q used to calculate $\langle dH(q)/dq\rangle_q$ yield a fractional electron localized on Ni. The exception is $q = 0.95$. Recall that $\langle dH(q)/dq\rangle_q$ for this value of q is generated using $q = 1$, which reflects the Ni^+ ion in water. This ion is stable, and no electron delocalization occurs. However, when snaphots along this $q = 1$ AIMD trajectory are taken and q is changed to 0.95 to perform finite difference calculations of $\langle dH(q)/dq\rangle_q$, maximally, Wannier function analysis reveals that the highest occupied molecular orbital containing the fractional (0.05) electron is now centered several Ångstroms away from the Ni nucleus. In other words, the $Ni^{+0.95}/32\ H_2O$ simulation actually represents a Ni^+ and an $e^{0.05-}$ delocalized away from the Ni. Therefore, $\langle dH(q)/dq\rangle_q$ for $(q = 1 - \delta q)$ should not be calculated directly, and instead might be extrapolated from smaller q values. Since $\langle dH(q)/dq\rangle_q$ already lies on a straight line with this end point q value, however, performing the extrapolation would not change $\Delta\Delta G_{hyd}$ significantly, and we have not pursued this avenue.

As discussed above, $Ni^{(0)}$ is strongly bonded to two H_2O molecules. This makes calculating the singlet $Ni^{(0)}$ ΔG_{hyd} using the AIMD method difficult because the reference system is the unsolvated triplet $Ni^{(0)}$. An integration path from that species to $Ni^{(0)}(H_2O)_2$ is not readily available. An attempt to use a coordination constraint reaction coordinate[94] to break $Ni-H_2O$ bonds fails to achieve a sufficient $Ni-O_w$ separation such that the subsequent spin-flip energy can be matched to the gas-phase value. As a result, we have relied on QCT to calculate the ΔG_{hyd} of $Ni^{(0)}$.

Recall that AIMD/PBE and AIMD/DFT+U trajectories are used to generate the snapshots where $\langle dH(q)/dq\rangle_q$ is computed for $q \leq 1$ and $q \geq 1$, respectively, using the DFT+U method. In Figure 5, the effect of using AIMD/DFT+U trajectories (instead of AIMD/PBE ones) is depicted for $q = 0$ and $q = 1$. Using configurations from the former functional yields average $\langle dH(q)/dq\rangle_q$ values that differ by -0.16 and $+0.04$ eV compared to AIMD/PBE configurations for these two q values. (On the scale of the graph, these small differences are almost indistinguishable.) $q = 0$ is expected to give the largest discrepancy because the DFT+U and PBE methods predict $Ni^{(0)}(H_2O)_2$ binding energies that differ the most (by 80%, section III.A). Nevertheless, the hydration *structures* obtained are sufficiently similar in that snapshots from either type of trajectory can be used to compute $\langle dH(q)/dq\rangle_q$. In the SI, we further show that the $g(r)$ values obtained in AIMD/DFT+U and AIMD/PBE trajectories are very similar. Since $\Delta\Delta G_{hyd}$ is obtained by integrating over the entire range of q, the average of these two values, -0.06 eV or -1.4 kcal/mol, can be taken as an estimate of the small discrepancy in $\Delta\Delta G_{hyd}$ one can expect if AIMD/DFT+U were used to generate trajectories throughout.

Finally, we consider the differential hydration free energy between Ni_2 and Ni_2^+ using the QCT method. Table 2 shows that the spin quartet yields $\Delta\Delta G_{hyd} = -4.42$ eV using the B3LYP functional. These changes in hydration free energies are critical for calculating redox potentials, described in the next subsection.

III.E. Redox Potentials. Estimating the Φ_{redox} of Ni and Ni_2 species requires ionization potentials in addition to $\Delta\Delta G_{hyd}$. As discussed in our previous work,[53,54] the first and second IP of Ni are not accurately calculated by either the PBE or B3LYP functional (see the SI). Instead, we adopt the widely accepted experimental values of 7.640 and 18.170 eV for the IPs.[95,96] Combining the IP and $\Delta\Delta G_{hyd}$, and subtracting the -4.44 V associated with SHE, the QCT/B3LYP Φ_{redox} for $Ni^{(0)} \rightarrow Ni^+$ becomes -2.18 V. The AIMD/DFT+U Φ_{redox} is a similar value, -2.17 V. These results are consistent with the view that excess electrons (-2.8 V), but not electron-scavenging organic radical anions (-1.2 V), can reduce Ni^+ to a neutral Ni atom dispersed in water.

In contrast, QCT/B3LYP yields $\Phi_{redox} = -1.05$ V for $Ni^+ \rightarrow Ni^{2+}$. This value is within the theoretical uncertainty of the hydroxymethyl radical anion $\Phi_{redox} = -1.18$ V. (Here, the uncertainty in the predicted value is estimated at $\sim$0.2 V, arising from the use of the B3LYP functional for hydration energy compared to CCSD(T).) This suggests that organic radical anions may already be able to reduce Ni^{2+} to Ni^+ in aqueous solutions. AIMD/DFT+U simulations yield a slightly more negative Φ_{redox} of -1.28 V, which is still reasonably close to the hydroxymethyl radical anion Φ_{redox}.

Our estimate of the Ni^{2+} reduction potential is consistent with the observation that Ni^+ does not appear to reduce Ag^+ to a Ag atom in water.[9] This suggests that the Ni^{2+} (aq) one-electron reduction reaction exhibits a Φ_{redox} less negative than the Ag^+ redox reaction reported at -1.75 V versus SHE.[9] Indeed, the QCT and AIMD estimates of Ni^{2+} Φ_{redox} significantly exceed -1.75 V. Furthermore, in the radiolysis-assisted synthesis of Ni/Pd nanocluster alloys,[8] the $Pd^{2+} \rightarrow Pd^+$ Φ_{redox} is known to exceed -1.18 V, meaning that Pd^{2+} is readily reduced to Pd^+ by organic electron scavengers. If $Ni^{2+} \rightarrow Ni^+$ indeed had a Φ_{redox} of -2.7 V,[14] much more negative than -1.18 V, nanoparticle alloy synthesis may seem more difficult to initiate because Pd might have been preferentially nucleated first. But NiPd alloys are indeed observed in experiments.[8] This observation is arguably more consistent with our revised estimate of Ni^{2+} Φ_{redox}. As discussed in section III.C, in arriving at their -2.7 V estimate for this reduction reaction,[13] Dixon and Baxendale adopted very different ΔG_{hyd} theoretical estimates. Our modern electronic structure calculations thus yield qualitatively different conclusions compared to earlier calculations reported in the radiolysis literature.

Table 3. Free Energies of Dimerization and Disproportionation Reactions[a]

reaction	method	functional	ΔG (eV)
$2Ni^{(0)} \rightleftharpoons Ni_2^{(0)}$	QCT	B3LYP	−1.45 eV
$Ni^+ + Ni^{(0)} \rightleftharpoons Ni_2^+$	QCT	B3LYP	−0.41 eV
$2Ni^+ \rightleftharpoons Ni^{(0)} + Ni^{2+}$	QCT	B3LYP	+1.42 eV
	QCT	B3LYP	+1.12 eV*
	AIMD	DFT+U	+1.17 eV*

[a] Asterisks indicate that the DFT IPs implicit in the reactions have been replaced by experimental values.

The experimental IP for Ni_2 exhibits extremely large uncertainties.[23] Fortunately, unlike the case of the Ni atom, we find that B3LYP and CCSD(T) yield IP values that are in reasonable agreement with each other (7.73 and 7.42 eV, respectively) when using the 6-311+G(d,p) basis set. This gives us the confidence to adopt the zero temperature CCSD(T) value of 7.42 eV as the N_2 IP. Combined with the $\Delta\Delta G_{hyd}$, the B3LYP quasi-chemical Φ_{redox} for reduction of the monovalent nickel dimer, Ni_2^+ becomes −1.44 V. This is lower than the monomeric Ni^+ reduction potential.

III.F. Dimerization and Disproportionation Reactions. Finally, we consider reactions that involve two Ni but no excess electrons (Table 3). Dimerization to form $Ni_2^{(0)}$ and Ni_2^+ from $Ni^{(0)}$ and Ni^+ are predicted to be favorable in liquid water. The Ni^+ disproportionation reaction into $Ni^{(0)}$ and Ni^{2+} is, however, unfavorable. The latter prediction conforms to at least one published experimental result in the literature where Ni atoms are not detected in the presence of Ni^+ (aq).[97] Ni_2^+ disproportionation into Ni^+ and $Ni^{(0)}$, the inverse of the second reaction in Table 3, is also unfavorable.

The AIMD results in Table 3 apply the hydration free energies listed in Table 2 and experimental ionization potentials, while the QCT approach implicitly uses B3LYP IP. If experimental IPs are adopted in QCT as well, better quantitative agreement between the AIMD and QCT values is obtained. The qualitative conclusion, however, remains the same: Ni^+ disproportionation remains strongly unfavorable.

Our results in sections III.E and III.F are consistent with the following overall picture. Ni^{2+} is reduced to Ni^+ by solvated electrons and possibly organic radical anions. Ni^+ does not disporportionate into $Ni^{(0)}$ and Ni^{2+}; it can, however, be reduced by hydrated excess electrons to $Ni^{(0)}$. The $Ni^{(0)}$'s are then immediately consumed to produce larger clusters. These seed clusters become the nuclei of Ni metal nanoparticles.

IV. Conclusions

We have applied first principles methods to calculate the redox potentials (Φ_{redox}) and dimerization free energies of monomeric and dimeric Ni^{q+} species in liquid water. AIMD/DFT and QCT/B3LYP yield Ni^{2+} (aq) one-electron reduction potentials of −1.05 and −1.28 V. These are at least 1.4 V less negative than previous estimates in the radiolysis literature.[13,14] Φ_{redox}'s predicted from our state-of-the-art electronic structure calculations are within the combined experimental and theoretical uncertainties of the redox potential associated with hydroxymethyl radical anions ($\Phi_{redox} = -1.18$ V). Our findings suggest that these electron-scavenging organic species, in addition to hydrated electrons (−2.7 to −2.9 V), may be able to reduce Ni^{2+} to Ni^+ in water. Ni^+ does not readily disproportionate into $Ni^{(0)}$ and Ni^{2+}.

Even though our calculations are limited to dimers, we can make the following mechanistic prediction. In the beginning, $Ni^+ + e^- \rightarrow Ni^{(0)}$ and $Ni^{2+} + e^- \rightarrow Ni^+$ readily occurs in γ-irradiated solutions. From the calculated redox potentials (Table 2), the reducing agent is the solvated excess electron in the former case, but the reduction of Ni^{2+} can be accomplished via organic radical anions. In the next step, the reactions $2Ni^{(0)} \rightarrow Ni_2^{(0)}$ and $Ni^+ + Ni^{(0)} \rightarrow Ni_2^+$ take place (Table 3, reactions 1 and 2). Even though $Ni^{(0)}$ can spontaneously undergo a disproportionation reaction with Ni^{2+} to form Ni^+ (Table 3, third reaction), some dimer formation should occur at a sufficient $Ni^{(0)}$ concentration (i.e., at high radiation dosage). We stress that we have elucidated the thermodynamic feasibility of the reactions but not their relative rates. In γ-radiolysis, the driving force is significant, and the overall reaction may be kinetically limited. These monomers and dimers serve as nuclei for metal nanoparticle growth, yielding larger clusters that coalesce from them. Theoretical studies on these larger clusters will be performed in the future.

Good agreement exists between the quasi-chemical (QCT) method using dielectric continuum approximation for outer-shell hydration contributions and *ab initio* molecular dynamics (AIMD) simulations where outer-shell H_2O molecules are explicitly present. Thus, this study confirms the advantage of using the QCT approach when dealing with transition metal ions with rigid hydration shells.

Acknowledgment. This work was supported by the Department of Energy under Contract DE-AC04-94AL85000, by Sandia's LDRD program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy.

Supporting Information Available: Further information is provided regarding ionization potentials, spin splittings, zero temperature binding energies of $Ni-H_2O$ complexes, zero point energy contributions, and more comparison between DFT+U and B3LYP predictions. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Bartolozzi, M. *J. Power Sources* **1989**, *27*, 219.

(2) Kongolo, K.; Mwema, M. D. *Hyperfine Interact.* **1998**, *111*, 281.

(3) Yamamoto, T.; Liimatainen, J.; Linden, J.; Karppinen, M.; Yamauchi, H. *J. Mater. Chem.* **2000**, *10*, 2342.

(4) Lovell, M. A.; Xie, C.; Xiong, S.; Markesbery, W. R. *J. Alzheimer's Disease* **2003**, *5*, 229.

(5) Desideri, F.; Polticelli, F.; Falconi, M.; Sette, M.; Ciriolo, M. R.; Paci, M.; Rotilio, G. *Arch. Biochem. Biophys.* **1993**, *301*, 244.

(6) Breitenkamp, M.; Henglein, A.; Lilie, J. *Ber. Bunsen. Phys. Chem.* **1976**, *80*, 973.

(7) Zhang, Z. Y.; Nenoff, T. M.; Huang, J.; Berry, D. T.; Provencio, P. *J. Phys. Chem. C* **2009**, *113*, 1155.

(8) Zhang, Z.; Nenoff, T. M.; Leung, K.; Ferreira, S. R.; Huang, J. Y.; Berry, D. T.; Provencio, P. P.; Stumpf, R. *J. Phys. Chem. C*. **2010**, *114*, 14309.

(9) Ershov, B. G.; Janata, E.; Henglein, A. *J. Phys. Chem.* **1994**, *98*, 7619.

(10) Belloni, J. *Catal. Today* **2006**, *113*, 141.

(11) Khatouri, J.; Mostavi, M.; Amblard, J.; Belloni, J. *Z. Phys. D* **1993**, *26*, S82.

(12) Gachard, E.; Remita, H.; Khatouri, J.; Keita, B.; Nadjo, L.; Belloni, J. *New J. Chem.* **1998**, 1257.

(13) Baxendale, J. H.; Dixon, R. S. *Z. Phys. Chem. (Munich)* **1964**, *43*, 161.

(14) Baxendale, J. H.; Keene, J. P.; Scott, D. A. *Chem. Commun.* **1966**, *20*, 715.

(15) Ershov, B. G. *Russ. Chem. Bull.* **1999**, *48*, 1.

(16) Jacobson, D. B.; Freiser, B. S. *J. Am. Chem. Soc.* **1986**, *108*, 27.

(17) Klotzbuecher, W.; Ozin, G. A. *Inorg. Chem.* **1977**, *16*, 984.

(18) Koretsky, G. M.; Kerns, K. P.; Nieman, G. C.; Knickelbein, M. B.; Riley, S. J. *J. Phys. Chem. A* **1999**, *103*, 1997.

(19) Sun, Y.; Fournier, R.; Zhang, M. *Phys. Rev. A* **2009**, *79*, 043202.

(20) Ershov, B. G.; Janata, E.; Henglein, A.; Fotjik, A. *J. Phys. Chem.* **1993**, *97*, 4589.

(21) Janata, E.; Henglein, A.; Ershov, B. G. *J. Phys. Chem.* **1994**, *98*, 10888.

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

(23) Rohlfing, E. A.; Cox, D. M.; Kaldor, A. *J. Phys. Chem.* **1984**, *88*, 4497.

(24) Michelini, M. C.; Diez, R. P.; Jubert, A. H. *Comput. Mater. Sci.* **2004**, *31*, 292.

(25) Merchan, M.; Pou-Amerigo, R.; Roos, B. O. *Chem. Phys. Lett.* **1996**, *252*, 405.

(26) Reddy, B. V.; Nayak, S. K.; Khanna, S. N.; Rao, B. K.; Jena, P. *J. Phys. Chem. A* **1998**, *102*, 1748.

(27) Andrés Cisneros, G.; Castro, M.; Salahub, D. R. *Int. J. Quantum Chem.* **1999**, *75*, 847.

(28) Knickelbein, M. B.; Yang, S.; Riley, S. J. *J. Chem. Phys.* **1990**, *93*, 94.

(29) Arvizu, G. L.; Calaminici, P. *J. Chem. Phys.* **2007**, *126*, 194102.

(30) Belloni, J.; Mostafavi, M.; Remita, H.; Marignier, J. L.; Delcourt, M. O. *New J. Chem.* **1998**, *22*, 1239.

(31) Henglein, A.; Meisel, D. *Langmuir* **1998**, *14*, 7392.

(32) Wardman, P. *J. Phys. Chem. Ref. Data* **1989**, *18*, 1637.

(33) Baxendale, J. H. *Radiat. Res. Suppl.* **1964**, *4*, 114.

(34) Noyes, R. M. *J. Am. Chem. Soc.* **1962**, *84*, 513.

(35) Asthagiri, D.; Pratt, L. R.; Paulaitis, M. E.; Rempe, S. B. *J. Am. Chem. Soc.* **2004**, *126*, 1285.

(36) Li, J.; Fisher, C. L.; Chen, J. L.; Bashford, D.; Noodleman, L. *Inorg. Chem.* **1996**, *35*, 4694.

(37) Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R. *J. Phys. Chem. A* **2009**, *113*, 6745.

(38) Galstyan, A.; Knapp, E. W. *J. Comput. Chem.* **2009**, *30*, 203.

(39) Tsushima, S. *J. Phys. Chem. B* **2008**, *112*, 13059.

(40) Uudsemaa, M.; Tamm, T. *J. Phys. Chem. A* **2003**, *107*, 9997.

(41) Jaque, P.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, *111*, 5783.

(42) Baik, M.; Friesner, R. A. *J. Phys. Chem. A* **2002**, *106*, 7407.

(43) Sabo, D.; Varma, S.; Martin, M. G.; Rempe, S. B. *J. Phys. Chem. B* **2008**, *112*, 867.

(44) Varma, S.; Rempe, S. B. *J. Am. Chem. Soc.* **2008**, *130*, 15405.

(45) Asthagiri, D.; Dixit, P. D.; Merchant, S.; Paulaitis, M.; Pratt, L. R.; Rempe, S. B.; Varma, S. *Chem. Phys. Lett.* **2010**, *485*, 1.

(46) Pratt, L. R.; LaViolette, R. A. *Mol. Phys.* **1998**, *94*, 909. Pratt, L. R.; Rempe, S. B. *Simulation and Theory of Electrostatic Interactions in Solution*; Pratt, L. R., Hummer, G., Eds.; AIP: New York, 1999; pp 172−201. Beck, T. L.; Paulaitis, M. E.; Pratt, L. R. *The Potential Distribution Theorem: Models of Molecular Solutions*; Cambridge University Press: New York, 2006; pp 166−195. Pratt, L. R.; Asthagiri, D. *Free Energy Calculations*; Chipot, C., Pohorille, A., Eds.; Springer-Verlag: Berlin, 2007; pp 323−352.

(47) See, e.g.: Swift, T. J.; Connick, R. E. *J. Chem. Phys.* **1962**, *37*, 307. The exchange of water molecules in the Ni^{2+} first hydration shell with bulk water was estimated to occur on 10 ps timescales. We did not observe such exchanges when generating the Ni^{2+} (aq) trajectory in ref 53. However, we adopted deuterium mass for protons and used a thermostat; therefore, the AIMD dynamics cannot be directly compared with experimental timescales.

(48) Blumberger, J.; Sprik, M. *J. Phys. Chem. B.* **2004**, *108*, 6529.

(49) Blumberger, J.; Tateyama, Y.; Sprik, M. *Comput. Phys. Commun.* **2005**, *169*, 256.

(50) Blumberger, J. *J. Am. Chem. Soc.* **2008**, *130*, 16065.

(51) VandeVondele, J.; Ayala, R.; Sulpizi, M.; Sprik, M. *J. Electroanal. Chem.* **2007**, *607*, 113. Tateyama, Y.; Blumberger, J.; Ohno, T.; Sprik, M. *J. Chem. Phys.* **2007**, *126*, 204506.

(52) Kollman, P. A. *Chem. Rev.* **1983**, *93*, 2395.

(53) Leung, K.; Rempe, S. B.; von Lilienfeld, O. A. *J. Chem. Phys.* **2009**, *130*, 204507.

(54) Rempe, S. B.; Leung, K. *J. Chem. Phys.* **2010**, *133*, 047104.

(55) Rempe, S. B.; Pratt, L. R.; Hummer, G.; Kress, J. D.; Martin, R. L.; Redondo, A. *J. Am. Chem. Soc.* **2000**, *122*, 966. Rempe, S. B.; Pratt, L. R. *Fluid Phase Equilib.* **2001**, *183*, 121. Rempe, S. B.; Asthagiri, D.; Pratt, L. R. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1966.

(56) Varma, S.; Rempe, S. B. *Biophys. J.* **2007**, *93*, 1093.

(57) Varma, S.; Sabo, D.; Rempe, S. B. *J. Mol. Biol.* **2008**, *376*, 13.

(58) Pasquarello, A.; Petri, I.; Salmon, P. S.; Parisel, O.; Car, R.; Toth, E.; Powell, D. H.; Fischer, H. E.; Helm, L.; Merbach, A. E. *Science* **2001**, *291*, 856.

(59) Bernasconi, L.; Blumberger, J.; Sprik, M.; Vuilleumier, R. *J. Chem. Phys.* **2004**, *121*, 11885. Sherman, D. M. *Geomchim. Cosmochim. Acta* **2007**, *71*, 714.

(60) Schwenk, C. F.; Rode, B. M. *Chem. Phys. Chem.* **2003**, *4*, 931.

(61) Schwenk, C. F.; Rode, B. M. *J. Chem. Phys.* **2003**, *119*, 9523. Blumberger, J.; Bernasconi, L.; Tavernelli, I.; Vuilleumier, R.; Sprik, M. *J. Am. Soc. Chem.* **2004**, *126*, 3928.

(62) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. *Phys. Rev. B* **1991**, *44*, 943. Liechtenstein, A. I.; Anisimov, V. I.; Zaanen, J. *Phys. Rev. B* **1995**, *52*, 5467.

(63) The DFT+U implementation in VASP was described in: Rohrbach, A.; Hafner, J.; Kresse, G. *Phys. Rev. B* **2004**, *69*, 075413.

(64) Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N. *Phys. Rev. Lett.* **2006**, *97*, 103001. Scherlis, D. A.; Cococcioni, M.; Sit, P. H. L.; Marzari, N. *J. Phys. Chem. B* **2007**, *111*, 7384. Sit, P. H. L.; Cococcioni, M.; Marzari, N. *J. Electroanal. Chem.* **2007**, *607*, 107. Leung, K.; Rempe, S. B.; Schultz, P. A.; Sproviero, E. M.; Batista, V. S.; Chandross, M. E.; Medforth, C. J. *J. Am. Chem. Soc.* **2006**, *128*, 3659. Leung, K.; Medforth, C. J. *J. Chem. Phys.* **2007**, *126*, 024501. Leung, K.; Nielsen, I. M. B.; Sai, N.; Medforth, C. J.; Shelnutt, J. A. *J. Phys. Chem. A*. **2010**, *114*, 10174. Panchmatia, P. M.; Sanyal, B.; Oppeneer, P. M. *Chem. Phys.* **2008**, *343*, 47. Oppeneer, P. M.; Panchmatia, P. M.; Sanyal, B.; Eriksson, O.; Ali, M. E. *Prog. Surf. Sci.* **2009**, *84*, 18.

(65) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C. Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian*; Gaussian, Inc.: Wallingford, CT, 2009.

(66) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968. Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *76*, 1910.

(67) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372. Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

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

(69) Stefanovich, E. V.; Truong, T. N. *Chem. Phys. Lett.* **1995**, *244*, 65.

(70) Ohtaki, H.; Radnai, T. *Chem. Rev.* **1993**, *93*, 1157.

(71) Bol, W.; Gerrits, G. J.; Panthale, C. L. *J. Appl. Crystallogr.* **1970**, *3*, 486.

(72) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10037.

(73) Bank, R.; Holst, M. *Soc. Ind. Appl. Math J. Sci. Comput.* **2000**, *22*, 1411.

(74) Kresse, G.; Furthmüller, J. *Phys. Rev. B.* **1996**, *54*, 11169. Ibid. *Comput. Mater. Sci.* **1996**, *6*, 15.

(75) Blöchl, P. E. *Phys. Rev. B* **1994**, *50*, 17953. 1994.

(76) The VASP implementation is discussed in: Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*, 1758.

(77) Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757.

(78) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388, and references therein.

(79) Schwegler, E.; Grossman, J. C.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *121*, 5400. Sit, P. H.-L.; Marzari, N. *J. Chem. Phys.* **2005**, *122*, 204510. Rempe, S. B.; Mattsson, T. R.; Leung, K. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4685.

(80) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.

(81) Nobuyuki, O.; Christopher, S.; Sapan, A. S.; Marcos, M. G.; Axel, T. B. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 641.

(82) Leung, K.; Marsman, M. *J. Chem. Phys.* **2007**, *127*, 154722.

(83) Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Phys. Chem.* **1996**, *100*, 1206.

(84) Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Chem. Phys.* **1997**, *107*, 9275.

(85) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847.

(86) Pratt, L. R. *J. Phys. Chem.* **1992**, *96*, 25. Wilson, M. A.; Pohorille, A.; Pratt, L. R. *J. Chem. Phys.* **1989**, *90*, 5211. Wilson, M. A.; Pohorille, A.; Pratt, L. R. *J. Phys. Chem.* **1987**, *91*, 4873. Sokhan, V. P.; Tildesley, D. J. *Mol. Phys.* **1997**, *it*, 625.

(87) Leung, K. *J. Phys. Chem. Lett.* **2010**, *1*, 496.

(88) Jackson, J. D. *Classical Electrodynamics*; Wiley: New York, 1999; Chapter 2.

(89) Orgel, L. E. *An Introduction to Transition-Metal Chemistry: Ligand-Field Theory*; Methuen & Co.: London, 1960; pp 46.

(90) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833. Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233.

(91) Zhang, F. S.; Lynden-Bell, R. M. *Phys. Rev. Lett.* **2003**, *90*, 185505.

(92) Marcus, Y. *Biophys. Chem.* **1994**, *51*, 111.

(93) See, e.g.: Mori-Sánchez, P.; Cohen, A. J.; Yang, W. T. *J. Chem. Phys.* **2006**, *125*, 201102.

(94) Sprik, M. *Chem. Phys.* **2000**, *258*, 139.

(95) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2006**, *125*, 074110.

(96) Shenstone, A. G. *J. Res. Natl. Bur. Stand. (U.S.)* **1970**, *74A*, 80.

(97) Kelm, M.; Lilie, J.; Henglein, A.; Janata, E. *J. Phys. Chem.* **1974**, *78*, 882.

CT100431M

JCTC Journal of Chemical Theory and Computation

Efficient Calculation of QM/MM Frequencies with the Mobile Block Hessian

An Ghysels,*[,†] H. Lee Woodcock III,*[,‡] Joseph D. Larkin,[§] Benjamin T. Miller,[§] Yihan Shao,[§,⊥] Jing Kong,[⊥] Dimitri Van Neck,[†] Veronique Van Speybroeck,[†] Michel Waroquier,[†] and Bernard R. Brooks[§]

Center for Molecular Modeling, Ghent University, Technologiepark 903, 9052 Zwijnaarde, Belgium, Department of Chemistry, University of South Florida, 4202 E. Fowler Avenue, CHE 205, Tampa, Florida 33620-5240, United States, Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, United States, and Q-Chem Inc., 5001 Baum Blvd, Suite 690, Pittsburgh, Pennsylvania 15213, United States

Received August 23, 2010

Abstract: The calculation of the analytical second derivative matrix (Hessian) is the bottleneck for vibrational analysis in QM/MM systems when an electrostatic embedding scheme is employed. Even with a small number of QM atoms in the system, the presence of MM atoms increases the computational cost dramatically: the long-range Coulomb interactions require that additional coupled perturbed self-consistent field (CPSCF) equations need to be solved for each MM atom displacement. This paper presents an extension to the Mobile Block Hessian (MBH) formalism for QM/MM calculations with blocks in the MM region and its implementation in a parallel version of the Q-Chem/CHARMM interface. MBH reduces both the CPU time and the memory requirements compared to the standard full Hessian QM/MM analysis, without the need to use a cutoff distance for the electrostatic interactions. Special attention is given to the treatment of link atoms which are usually present when the QM/MM border cuts through a covalent bond. Computational efficiency improvements are highlighted using a reduced chorismate mutase enzyme system, consisting of 24 QM atoms and 306 MM atoms, as a test example. In addition, the drug bortezomib, used for cancer treatment of myeloma, has been studied as a test case with multiple MBH block choices and both a QM and QM/MM description. The accuracy of the calculated Hessians is quantified by imposing Eckart constraints, which allows for the assessment of numerical errors in second derivative procedures. The results show that MBH within the QM/MM description not only is a computationally attractive method but also produces accurate results.

I. Introduction

Normal mode analysis (NMA) is a well-known technique which estimates the intrinsic vibrational frequencies of chemical systems by assuming a harmonic shape for the potential energy surface. Despite its simplicity, it is still a popular and effective approach for predicting vibrational IR and Raman spectra,[1] for identifying chemical groups,[2] or for studying the large-amplitude collective motions involved in conformational changes of biomolecules.[3] This method is based on the diagonalization of the Hessian matrix, which contains the second derivatives of the potential energy with respect to the nuclear coordinates. The calculation of this $3N_{AT} \times 3N_{AT}$ Hessian, where N_{AT} is the number of atoms,

* Corresponding author e-mail: an.ghysels@ugent.be (A.G.), hlw@mail.usf.edu (H.L.W.).
† Ghent University.
‡ University of South Florida.
§ National Institutes of Health.
⊥ Q-Chem Inc.

10.1021/ct100473f
Published on Web 01/06/2011

is computationally expensive when a quantum mechanical (QM) description is used, because a set of $3N_{AT}$ coupled perturbed self-consistent field (CPSCF) equations needs to be solved.[4−11] In contrast, the computational load of the second derivative calculation is in comparison extremely cheap in the molecular mechanics (MM) description, but force fields cannot, in general, be used to investigate chemical reactions where the change in electron density (i.e., bond making/breaking, radical processes, etc.) is a purely quantum mechanical phenomenon.

Hybrid QM/MM models aim at combining the best features of the QM and MM models: the quantum descriptions necessary for chemistry and the computational advantages of force fields. The QM/MM approach partitions the system into a QM region for the chemically interesting site and an MM region for the surrounding chemical environment.[12−15] The effective cost of a QM/MM Hessian calculation depends heavily on the treatment of the electrostatics between the QM and MM region, for which two schemes have been developed.

In a subtractive scheme like the original version of ONIOM, the potential energy is the sum of the MM energy of the whole system, plus the QM energy of the QM region, minus the MM energy of the QM region as a correction for double counting.[16,17] During the QM part of the calculation, the QM atoms are unaware of the existence of the MM atoms, and thus the electron cloud in the QM region is not influenced by the MM partial charges, i.e., mechanical embedding. As a consequence, the displacement of an MM atom does not cause a change in the QM wave function, such that the corresponding derivatives are simply equal to zero, and there is no need to solve CPSCF equations for MM atom displacements.

In an additive scheme, however, as implemented in the Q-Chem/CHARMM interface[18−20] and in many other interfaces, the potential energy consists of the QM energy of the QM atoms, the MM energy of the MM atoms, and the Coulomb and van der Waals interaction energy between QM and MM atoms.[12,21,22] Such a description provides a more accurate treatment of the long-range electrostatics, which is invaluable when studying, for example, reactions and molecular configurations. This idea has also been applied to the original ONIOM scheme to account for the polarization effects from the MM region.[23] In the additive scheme, every displacement of an MM atom leads to an additional CPSCF equation.[24] Even when the number of QM atoms is low, the QM/MM interaction term in the Hamiltonian makes the Hessian determination too costly for systems with a large number of MM atoms.

Recently, the mobile block Hessian approach was developed[25−27] to calculate frequencies in a partially optimized structure. The method groups atoms into blocks which are restricted to rigid motions during the vibrational analysis. The internal geometry is fixed, but each block is still allowed to translate or rotate as a whole. Consequently, block motions replace the individual atom motions in the CPSCF equations, thus reducing the number of CPSCF equations. Until now, no implementation was available that exploits the computational advantage offered by solving CPSCF equations of reduced dimension. In this paper, we present such an implementation for the specific case of blocks chosen in the MM region.

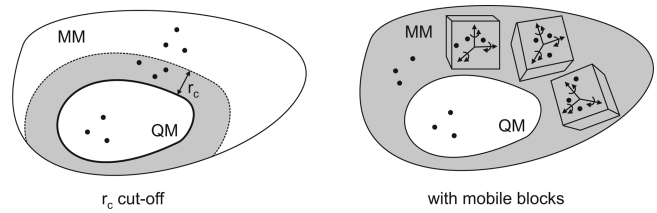


Figure 1. Cut-off technique versus mobile block Hessian approach. (a) When using a cutoff r_c to decrease the computational cost of the CPSCF, the electrostatic interaction between QM and MM atoms is neglected beyond the distance r_c. MM atoms outside the gray region do not interact with the QM atoms. (b) In the mobile block Hessian approach, all electrostatic interaction is still present. Blocks in the MM region are restricted to rigid body motions, i.e. translations and rotations of each block.

An alternative approach to economize on the number of CPSCF equations is the use of a cutoff distance r_c beyond which electrostatics between QM and MM atoms are neglected. Figure 1 illustrates the influence of a cutoff: the electrons in the QM region do not interact (Coulomb) with MM partial charges outside the gray cutoff zone. The MBH has the advantage that it does not neglect electrostatic interactions but rather restricts motions: the gray zone for MBH encompasses the whole MM region in Figure 1. Restricting the motion of distant blocks may change the overall vibrational free energy, but such effects are expected to largely cancel out when treated consistently in a thermodynamic cycle. It is the aim of this paper to show that the combination of MBH with the QM/MM description is a highly accurate and efficient approach. It therefore becomes the ideal alternative to the standard full Hessian calculation when the latter is no longer feasible. The Q-Chem/CHARMM interface now has a working parallel version for both the full QM/MM Hessian and mobile block QM/MM Hessian calculation.[28]

The following section presents the theoretical background on which the idea of MBH in a QM/MM description is based. First, the NMA equations, the frequency calculation in QM/MM, and the MBH equations are reviewed. Second, the adaptation needed for an efficient mobile block Hessian computation is outlined. Moreover, the treatment of multiple link atoms is clarified when the QM/MM border cuts through covalent bonds. It is also pointed out that MBH preserves the long-range electrostatic interactions, in contrast to the alternative approach with a cutoff distance r_c which induces an error decaying as slowly as $1/r_c^3$. The third section presents computational results of the chorismate mutase enzyme. This test case illustrates how MBH and parallellization reduce the memory requirements and CPU timings. In the fourth section, the oxidation of the bortezomib molecule is treated as a test case for the newly implemented method. This drug is used in cancer treatment since it inhibits the function of proteasomes upon binding, ultimately leading to cell death.[29,30] By imposing the Eckart constraints, the accuracy of the calculated Hessians is estimated. This

accuracy is then compared with the influence of various frequency treatments on the vibrational free energy differences: QM versus QM/MM and full Hessian versus MBH.

II. Theory and Implementation

II.A. Normal Mode Analysis. Assume that the positions of the N_{AT} atoms are described by Cartesian displacement coordinates, labeled $x = 1, ..., 3N_{AT}$, all with respect to a reference structure. A second order approximation of the potential energy surface around the reference structure is then equal to

$$V(x) \approx V(0) + G^T x + \frac{1}{2} x^T H x \tag{1}$$

where the reference energy $V(0)$ can be set to be zero. The $3N_{AT}$ dimensional gradient vector G contains the first derivatives, and the Hessian H is the $3N_{AT} \times 3N_{AT}$ matrix containing the second derivatives evaluated at the reference point. When calculating normal modes, the reference structure should be a stationary point on the potential energy surface, i.e. $G = 0$. Introducing the diagonal mass matrix M with the atomic masses on the diagonal, the normal-mode analysis (NMA) equations read

$$Hv = \omega^2 M v \tag{2}$$

Solving the NMA equations yields the eigenvalues ω^2 (frequency is $\nu = \omega/2\pi$) corresponding to the eigenvectors v.

Six frequencies should be zero because of the translational and rotational invariance of an isolated gas molecule (five for linear molecules). The corresponding normal modes represent global translations and rotations of the complete molecular system. It is possible to project out those zero frequency vectors before diagonalization, since their exact format is known. This projection amounts to imposing the Eckart constraints[26,31] and guarantees the presence of six frequencies that are identically zero even when the system is not perfectly at the stationary point on the energy surface or when the Hessian elements are inaccurate. The effect of the Eckart constraints on the frequencies is studied for the bortezomib example in section IV.

II.B. The QM/MM Full Hessian. In the additive scheme with electrostatic embedding, the system is separated into a QM and an MM region.[15] The QM region consists of N_{QM} nuclei and N_e electrons, described quantum mechanically in the Born−Oppenheimer approximation. The MM region contains N_{MM} partial charges, described classically, with $N_{QM} + N_{MM} = N_{AT}$. The Hamiltonian of the system of interacting QM and MM particles is written as

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}^{QM} + \hat{\mathcal{H}}^{QM/MM} + \mathcal{H}^{MM} \tag{3}$$

where $\hat{\mathcal{H}}^{QM}$ represents the Hamiltonian describing the QM region, i.e., the electronic kinetic energy and all electrostatic potentials generated by the electrons and QM nuclei. The QM/MM interaction Hamiltonian is

$$\hat{\mathcal{H}}^{QM/MM} = -\sum_{i=1}^{N_e}\sum_{k=1}^{N_{MM}} \frac{q_k}{|\boldsymbol{r}_i - \boldsymbol{R}_k|} + \sum_{n=1}^{N_{QM}}\sum_{k=1}^{N_{MM}} \frac{q_k Z_n}{|\boldsymbol{R}_n - \boldsymbol{R}_k|} + \sum_{n=1}^{N_{QM}}\sum_{k=1}^{N_{MM}} \left(\frac{A_{n,k}}{|\boldsymbol{R}_n - \boldsymbol{R}_k|^6} - \frac{B_{n,k}}{|\boldsymbol{R}_n - \boldsymbol{R}_k|^{12}} \right) \tag{4}$$

with q_k being the partial charge of the MM atom at position $\boldsymbol{R}_k$, Z_n the nuclear charge of the QM atom at position $\boldsymbol{R}_n$, $\boldsymbol{r}_i$ the electron positions, and $A_{n,k}$ and $B_{n,k}$ the van der Waals parameters. The QM/MM interaction Hamiltonian consists of the Coulomb interaction between MM charges and QM electrons, the Coulomb interaction between MM charges and QM nuclei, and the van der Waals interaction between MM atoms and QM atoms. The total energy of the system is thus given by

$$\begin{aligned} E_{tot} &= \langle \Phi | \hat{\mathcal{H}}^{QM} - \sum_{i=1}^{N_e}\sum_{k=1}^{N_{MM}} \frac{q_k}{|\boldsymbol{r}_i - \boldsymbol{R}_k|} | \Phi \rangle \\ &\quad + E_{nuc}^{QM/MM} + E_{vdW}^{QM/MM} + E^{MM} \\ &= E_{el}^{QM} + E_{el}^{QM/MM} + E_{nuc}^{QM/MM} + E_{vdW}^{QM/MM} + E^{MM} \end{aligned} \tag{5}$$

where $|\Phi\rangle$ is the electronic wave function for the QM atoms. The two electronic terms are calculated quantum mechanically and are referred to as the quantum part (denoted "quant"). The remaining three terms in this expression are classical (denoted "class"). This defines our decomposition of the total energy in a quantum and classical part:

$$E_{tot} = E_{quant} + E_{class} \tag{6}$$

The Cartesian Hessian H expresses the response of the total energy to $3N_{AT}$ Cartesian displacements. A general Hessian element is denoted as $H_{xy} = E_{tot}^{xy} = \partial^2 E_{tot}/\partial x \partial y$ ($x, y = 1, ..., 3N_{AT}$), with the superscripts referring to derivatives. The Hessian can be divided into submatrices as shown in Figure 2, depending on whether the x, y indices correspond to the QM−QM, MM−MM, or mixed QM−MM displacements. Not all terms in eq 5 contribute to each subblock of the Hessian: the derivatives of $E_{el,\,QM}$ and E_{MM} only contribute to the QM or MM subblock, respectively. However, the QM/MM interaction terms contribute to all subblocks of the Hessian. While the derivatives of the classical terms $E_{nuc,\,QM/MM}$ and $E_{vdW,\,QM/MM}$ are relatively easy to evaluate, the $E_{el,\,QM/MM}$ derivatives dominate the cost of the Hessian evaluation. Even if the number of QM atoms is low, the cost of the calculation still scales with N_{AT}. The main reason is that each MM atom adds three perturbations to the CPSCF equations, because the displacement of an external charge (an MM atom) leads to a change in the electronic wave function and the charge distribution. Section II.F shows how the number of perturbations can be reduced by the introduction of mobile blocks.

Methods based on the variational principle, such as Hartree-Fock or Kohn−Sham DFT, have the advantage that the $(2n + 1)$th derivative of the energy can be constructed from the nth derivative of the variational parameters (Wigner's $2n + 1$ theorem[32]). The variational

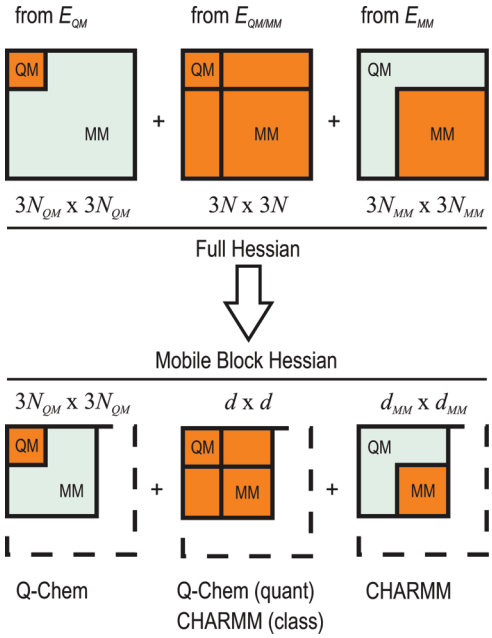


Figure 2. $\hat{\mathcal{H}}_{QM}$ and $\hat{\mathcal{H}}_{MM}$ contribute to the derivatives with respect to N_{QM} QM and N_{MM} MM atoms, respectively, whereas $\hat{\mathcal{H}}_{QM/MM}$ contributes to all Hessian elements. The mobile block Hessian H^{mb}, with blocks chosen in the MM region, is smaller in size ($d \times d$, $d = 3N_{QM} + d_{MM}$, bottom three panels) than the full Hessian H ($3N_{AT} \times 3N_{AT}$, upper three panels).

parameters are the elements of Θ that describe a unitary rotation among the molecular orbitals.[33] The parameters Θ are updated iteratively until the corresponding molecular orbitals are eigenfunctions of, e.g., the Fock operator in the case of Hartree−Fock-based methods. In accordance with Wigner's $2n + 1$ theorem, only the first derivatives of the Θ are needed for the construction of the second derivatives of the energy.

The specific discussion below applies to the Q-Chem/CHARMM interface, but in principle each QM and MM package with a suitable interface can be used for the construction of the full QM/MM Hessian. One should be aware that the implementation details may differ slightly depending on the choice of the QM and MM code. The implementation of the full QM/MM Hessian starts with the construction of the quantum contribution by the QM code. Using a compact notation,[34,35] where $\langle ... \rangle$ denotes the trace of a matrix, the Hartree−Fock energy E (corresponding to the first two terms in eq 5) calculated by the QM code reads

$$E = \langle PH_{\text{core}} \rangle + \frac{1}{2}\langle P\Pi P \rangle + \gamma = E(H_{\text{core}}, \Pi, S, \Theta) \quad (7)$$

where P ($= P(\Theta, S)$) is the density matrix, H_{core} is the core Hamiltonian matrix (including the Coulomb potential due to the MM partial charges, which corresponds to the first term in eq 4), Π represents the antisymmetrized two-electron integrals over spin orbitals, S is the overlap matrix, and γ is the nuclear repulsion energy of the QM atoms. The Fock operator is then defined as

$$F = H_{\text{core}} + P\Pi \quad (8)$$

and the standard self-consistent field (SCF) convergence criterion $E^{\Theta} = 0$ reads

$$FPS = SPF \quad (9)$$

The second derivatives with respect to atomic displacements are given by[11,33,34,36]

$$\frac{\partial^2 E}{\partial x \partial y} = E^{xy} = \langle PH^{xy}_{\text{core}} \rangle + \frac{1}{2}\langle P\Pi^{xy}P \rangle + \langle P^x H^y_{\text{core}} \rangle + \langle P^x \Pi^y P \rangle - \langle (PFP)S^{xy} \rangle - \langle (PFP)^x S^y \rangle + \gamma^{xy} \quad (10)$$

and require the density matrix response P^x, which is obtained by solving the coupled perturbated self-consistent field (CPSCF) equations for Θ^x. Since the energy is obtained by a variational method, the CPSCF equations can be derived from the identity $E^{\Theta} \equiv 0$:

$$(E^{\Theta})^x = E^{\Theta\Theta}\Theta^x + E^{\Theta H}H^x_{\text{core}} + E^{\Theta\Pi}\Pi^x + E^{\Theta S}S^x \equiv 0 \quad (11)$$

The calculation of the derivatives takes five steps:

(1) Construct $E^{\Theta H}H^x_{\text{core}} + E^{\Theta\Pi}\Pi^x + E^{\Theta S}S^x$.
(2) Solve CPSCF eq 11 for Θ^x.
(3) From Θ^x, construct P^x.
(4) From P^x, construct $F^x = H^x_{\text{core}} + \Pi^x P + \Pi P^x$.
(5) Construct E^{xy} according to eq 10.

As a result of the explicit QM/MM polarization effects, the CPSCF equations include perturbations for each MM atom, which makes step 2 and step 4 the most demanding. One should also pay attention to the memory requirements, which peak in step 2 and 4 because they scale as $6N_{MM}n_b^2$, where n_b is the number of basis functions in the basis set. To make the calculation more efficient, we have parallellized the quantum mechanical part (contribution from E^{QM}_{el} and $E^{QM/MM}_{el}$) of the full Hessian calculation. Section III discusses the timings and memory estimates in more detail using the chorismate mutase enzyme as a test system.

In the practical implementation, the next step involves passing the quantum mechanical information from Q-Chem back to CHARMM where the remaining classical terms are constructed and added to the Hessian at a relatively insignificant cost. The full QM/MM Hessian is then mass-weighted and diagonalized to obtain the frequencies and modes. The above discussion holds for Hartree−Fock calculations. The implementation for DFT is similar. The main difference is that for DFT one more term should be added to steps 1, 4, and 5 to account for the derivatives involving the exchange-correlation functional.

II.C. MBH Theory. The MBH method partitions the system into blocks of atoms.[25,26] During the geometry optimization, the position and orientation of each block are optimized, while the internal geometry of the blocks is not necessarily optimized. As a result of this partial optimization, there might be residual forces between the atoms within a block. Whereas in the subsequent vibrational analysis spurious imaginary frequencies might appear when applying the standard full Hessian NMA, the MBH is capable of reproducing physical frequencies.[25,26] The internal coordinates within multi-atom blocks are kept fixed, such that the

MBH model considers only a subset of the degrees of freedom (d). A single-atom block (free atom) is still described by its three Cartesian displacements, while a multi-atom block needs six parameters to describe the position and orientation (linear blocks are not considered here). Within our current implementation, all QM atoms are free, while MM atoms can either be free or become part of a multi-atom block. If d_b denotes the number of parameters of block b and N_{BL} is the number of blocks, the total number of parameters is given by

$$d = \sum_{b=1}^{N_{\mathrm{BL}}} d_b \tag{12}$$

with $d_b = 3$ for a single-atom block or $d_b = 6$ for a nonlinear block. A suitable choice of the block parameters is the convention introduced in ref 26. The parameters p_α, $\alpha = 1, ..., 6$, of a particular block give the position $\boldsymbol{r}'$ of each atom of the block with respect to the reference geometry $\boldsymbol{r}$ by successive rotations around the fixed z, y, and x axes of a space-fixed frame (p_6, p_5, p_4), followed by a translation (p_1, p_2, p_3):

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos p_4 & -\sin p_4 \\ 0 & \sin p_4 & \cos p_4 \end{pmatrix} \begin{pmatrix} \cos p_5 & 0 & \sin p_5 \\ 0 & 1 & 0 \\ -\sin p_5 & 0 & \cos p_5 \end{pmatrix} \times \begin{pmatrix} \cos p_6 & -\sin p_6 & 0 \\ \sin p_6 & \cos p_6 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{13}$$

The position of a single-atom block only needs the parameters that describe translation. In the following, we refer to the new set of dynamical variables with indices $p = 1, ..., d$. Useful quantities are the first (Jacobian) and second derivatives of the transformation between the Cartesian displacement coordinates and the block parameters, evaluated at the reference geometry:

$$T_{xp} = \frac{\partial x}{\partial p}; \quad C^{(x)}_{pp'} = \frac{\partial^2 x}{\partial p \partial p'} \tag{14}$$

The explicit expressions for the transformation matrices T (dimension $3N_{\mathrm{AT}} \times d$) and $C^{(x)}$, $x = 1, ..., 3N_{\mathrm{AT}}$ (each of dimension $d \times d$) have been derived in refs 26 and 37.

The second derivatives evaluated at the reference point with respect to the set of d parameters yield the mobile block Hessian H^{mb}, whose elements are defined as

$$H^{\mathrm{mb}}_{pp'} = \frac{\partial^2 E_{\mathrm{tot}}}{\partial p \partial p'} = E^{pp'}_{\mathrm{tot}} \tag{15}$$

The MBH elements are now related to the full Cartesian Hessian by the following transform

$$H^{\mathrm{mb}} = T^T H T + \sum_{x=1}^{3N_{\mathrm{AT}}} G_x C^{(x)} \tag{16}$$

and similarly for the gradient

$$G^{\mathrm{mb}} = T^T G = 0 \tag{17}$$

Note that, in the case of a partially optimized structure, the Cartesian gradient G might differ from zero, but the MB gradient G^{mb} should be zero because all block parameters are supposed to be optimized.

II.D. The QM/MM Mobile Block Hessian. By combining the QM/MM description with the mobile block concept, a considerable reduction in memory and CPU time becomes possible. This section explains the modifications of the CPSCF and the changes to the QM/MM interface that are required to realize the computational profit.

The decomposition in the classical and quantum terms of the total energy (see eq 6) leads to a similar decomposition of the gradient and the Hessian:

$$G = G_{\mathrm{class}} + G_{\mathrm{quant}} \tag{18}$$

$$H = H_{\mathrm{class}} + H_{\mathrm{quant}} \tag{19}$$

where, for instance, $G_{\mathrm{class},x} = \partial E_{\mathrm{class}}/\partial x$ and similar expressions hold for G_{quant}, H_{class}, and H_{quant}. To obtain the mobile block Hessian from the standard Cartesian Hessian calculation, the same decomposition is applied to eq 16:

$$\begin{aligned} H^{\mathrm{mb}} &= H^{\mathrm{mb}}_{\mathrm{class}} + H^{\mathrm{mb}}_{\mathrm{quant}} \\ &= T^T(H_{\mathrm{class}} + H_{\mathrm{quant}})T + \sum_{x=1}^{3N_{\mathrm{AT}}} (G_{\mathrm{class},x} + G_{\mathrm{quant},x})C^{(x)} \end{aligned} \tag{20}$$

As already mentioned, the classical part of the Hessian is computationally less demanding, but the quantum part of the Hessian is expensive. Instead of constructing the full quantum Hessian H_{quant} and then projecting it to the smaller mobile block Hessian dimension (eq 16), a substantial reduction in CPU time is obtained by directly constructing the mobile block Hessian for the quantum part, $H^{\mathrm{mb}}_{\mathrm{quant}}$. The elements are similar to those in eq 10, with block displacements p as perturbations instead of Cartesian displacements x. The density matrix response P^p is obtained by solving adapted CPSCF equations, $(E^{\Theta})^p = \Sigma_x (E^{\Theta})^x T_{xp} \equiv 0$. Similar to the full Hessian calculation, the construction of the derivatives occurs in five steps. However, at the end of step 1, the terms are projected with the T transform of eq 14, such that Θ^p can be solved from the adapted CPSCF:

$$E^{\Theta\Theta}\Theta^p = -\sum_x (E^{\Theta H} H^x_{\mathrm{core}} + E^{\Theta\Pi}\Pi^x + E^{\Theta S} S^x) T_{xp} \tag{21}$$

When the p index denotes a parameter of a block in the MM region, the summation over x is reduced to the Cartesian displacements of the MM atoms within the block only, such that the right-hand side of the transform greatly simplifies to

$$E^{\Theta\Theta}\Theta^p = -\sum_x E^{\Theta H} H^x_{\mathrm{core}} T_{xp} \tag{22}$$

After the transform to p variables, the rest of the steps are all very similar with x replaced by p.

In the remainder of this section, it is discussed how the QM code interacts with the MM code; it is, for instance, essential to add and project the matrices in the correct order.

The QM code returns a matrix of dimension $d \times d$, with d the total number of perturbations. When assembling the Hessian in step 5, the outcome is not the MB Hessian $H_{\text{quant}}^{\text{mb}}$ as in eq 20, but only the bilinear part

$$T^T H_{\text{quant}} T = H_{\text{quant}}^{\text{mb,bil}} \tag{23}$$

is obtained by construction. The gradient correction is still lacking but will be added at the end (see further, eq 26). To add the QM code contribution correctly to the MM code contribution, the Q-Chem Hessian $H_{\text{quant}}^{\text{mb, bil}}$ is first transformed to a matrix of the standard size $3N_{\text{AT}} \times 3N_{\text{AT}}$ by a linear transform Q of dimension $d \times 3N_{\text{AT}}$:

$$H'^{\text{bil}}_{\text{quant}} = Q^T H_{\text{quant}}^{\text{mb,bil}} Q \tag{24}$$

Here, Q is the pseudoinverse of the rectangular transform matrix T, such that $QT = 1$ and

$$T^T H'^{\text{bil}}_{\text{quant}} T = T^T Q^T H_{\text{quant}}^{\text{mb,bil}} QT = H_{\text{quant}}^{\text{mb,bil}} \equiv T^T H_{\text{quant}} T \tag{25}$$

The final expression for the mobile block Hessian reads

$$H^{\text{mb}} = T^T H_{\text{class}} T + T^T Q^T H_{\text{quant}}^{\text{mb,bil}} QT + \sum_{x=1}^{3N_{AT}} (G_{\text{class},x} + G_{\text{quant},x}) C^{(x)} \tag{26}$$

where the QM code (Q-Chem) calculates $H_{\text{quant}}^{\text{mb, bil}}$ and G_{quant} and performs the Q transform, while the MM code (CHARMM) calculates H_{class} and G_{class}, assembles classical and quantum parts, and performs the T transform and the gradient correction. In practice, the $E_{\text{nuc}}^{\text{QM/MM}}$ of eq 5 is calculated by Q-Chem, but in general, this is a purely classical term which could be calculated by either the MM or QM code.

The set of linear equations to be solved in step 2 (see eq 21) now counts d equations instead of $3N_{\text{AT}}$, leading to a computational profit especially on the level of memory requirements. Moreover, the code implemented in Q-Chem also works in parallel. Using N_{P} processors, the peak memory for MBH is equal to $2d_{\text{MM}}{n_{\text{b}}}^2/N_{\text{P}}$, where $d_{\text{MM}} = 6N_{\text{BL}} + 3N_{\text{MM, free}}$ is the number of MM perturbations, compared to the peak memory $6N_{\text{MM}}{n_{\text{b}}}^2/N_{\text{P}}$ for a full QM/MM Hessian calculation. The main difference with respect to the full Hessian calculation is that the factor $3N_{\text{MM}}$ is reduced to d_{MM}, which will generally be significantly smaller. The MBH also reduces the CPU time since fewer CPSCF equations need to be constructed and solved. This means that MBH allows vibrational analysis to be performed on larger systems than is feasible with the full Hessian. In section III, concrete timings and memory estimates are discussed for the chorismate mutase test system.

II.E. Treatment of Link Atoms. It is a common practice to introduce at least one link atom when cutting a bond across the QM/MM boundaries in the electrostatical embedding scheme.[15,38] To make the MBH available for a broader range of applications, our implementation has been extended to be able to treat link atoms correctly. The introduction of a link atom generates three additional degrees of freedom, leading to an extended potential energy surface $\tilde{V}(\boldsymbol{R}, \boldsymbol{R}_{\text{LK}})$. Therefore, the full Hessian $\tilde{H}$ of a system with N_{LK} link atoms has $3N_{\text{LK}}$ extra rows/columns. Diagonalization of this extended Hessian yields $3N_{\text{LK}}$ extra frequencies, which are in essence an artifact of the QM/MM border description and are not inherent to the real physical system. Moreover, the link atoms are usually not completely optimized during the energy minimization process, and unphysical imaginary frequencies might appear as a system with constrained link atoms in a nonequilibrium state.

Hence, one has to project out the $3N_{\text{LK}}$ link atom degrees of freedom to construct a Hessian H of the dimension $3N_{\text{AT}}$. A straightforward and simple solution is to omit the rows/columns in the Hessian that correspond to the link atoms.[39] This approach coincides with the Partial Hessian Vibrational Analysis (PHVA), which can be interpreted as associating an infinite mass to the link atoms.[39−45] This procedure not only disturbs the global translational and rotational symmetry of the system, reflected by the destruction of the six zero eigenvalues of the Hessian, but also the lower frequency spectrum is affected in an unpredictable manner. Cui and Karplus propose to project out the link atom motions after making them orthogonal to the global translation/rotation vectors.[24] This method could only be applied to a system containing a single link atom which is left unconstrained during the geometry optimization. This is often not the case, because preferably constraints on the link atom's position are used to locate it between the QM host and MM host. Such constraints ensure that one of the orbitals of the QM host is effectively pointed toward the MM host, thus providing a better description of the covalent bond.

Here, we focus on a different optimization procedure, which has also been investigated in the framework of ONIOM by Dapprich et al.[17,46] A similar methodology is now extended to the framework of QM/MM where explicit QM/MM polarization effects are included. Instead of a full geometry optimization, the position of the link atom is constrained and can be written as a function of the other QM and MM atom positions: $\boldsymbol{R}_{\text{LK}} = \boldsymbol{R}_{\text{LK}}(\boldsymbol{R})$. Respecting the following notation, where x stands for the $3N_{\text{AT}}$ displacements of the QM and MM atoms and x'' for the $3N_{\text{LK}}$ link atom displacements, we can express the constraints as $x'' = x''(x)$. They reduce the dimensionality of the potential energy surface $\tilde{V}$ to a new potential energy function in $3N_{\text{AT}}$ space

$$V(x) = \tilde{V}(x, x'')|_{x''=x''(x)} \tag{27}$$

Using the chain rule, the $3N_{\text{AT}}$-dimensional gradient G and $3N_{\text{AT}} \times 3N_{\text{AT}}$ Hessian H of the new function $V(x)$ can be written as

$$G^x = \tilde{G}^x + \sum_{x''} \tilde{G}^{x''} \frac{\partial x''}{\partial x} \tag{28}$$

$$\begin{aligned} H^{xy} = \; & \tilde{H}^{xy} + \sum_{x''} \tilde{H}^{x''y} \frac{\partial x''}{\partial x} + \sum_{y''} \tilde{H}^{xy''} \frac{\partial y''}{\partial y} \\ & + \sum_{x''y''} \tilde{H}^{x''y''} \frac{\partial x''}{\partial x} \frac{\partial y''}{\partial y} + \sum_{x''} \tilde{G}^{x''} \frac{\partial^2 x''}{\partial x \partial y} \end{aligned} \tag{29}$$

with $\tilde{G}$ being the original $(3N_{\text{AT}} + 3N_{\text{LK}})$-dimensional gradient vector of $\tilde{V}(x, x'')$ and $\tilde{H}$ the $(3N_{\text{AT}} + 3N_{\text{LK}}) \times (3N_{\text{AT}} + 3N_{\text{LK}})$ matrix containing the second derivatives of $\tilde{V}(x, x'')$,

all evaluated at the reference point. The equations for the Hessian elements (eq 29) indicate that the elimination of the link atom involves both projections of the original Hessian $\tilde{H}$ (first four terms on the right-hand side) as well as a term depending on the original forces $\tilde{G}$ on the link atom (last term on the right-hand side). The main point is that the constraints imposed during the geometry optimization are also imposed during the vibrational analysis, which is the key condition for consistent and meaningful frequencies.[26] Typically, the constraints $x''(x)$ only depend on the position of the neighboring QM host (x_{QMH}) and MM host (x_{MMH}). The constraint derivatives of type $\partial x''/\partial x$ or $\partial^2 x''/\partial x \partial y$ evaluated at the reference geometry are then only nonzero if x, y corresponds to host atom displacements. Consequently, only the rows/columns in the Hessian that involve host atoms are affected by the projection, while other gradient and Hessian elements remain unchanged, i.e., $\tilde{G}^x = G^x$ and $\tilde{H}^{xy} = H^{xy}$ if x, y do not involve host atom displacements.

In our QM/MM procedure, the link atom is forced to stay colinear with the QM and MM hosts during the energy minimization, and at a fixed, scaled distance. This completely determines the positions of the link atoms. In the subsequent vibrational analysis, one can either perform a numerical or an analytical second derivative calculation. For second derivatives obtained with numerical differentiation of slightly displaced geometries, the displaced geometries are such that they respect the constraints. Since the link atom degrees of freedom are thus never sampled, the numerical Hessian yields $3N_{LK}$ zero eigenvalues. For analytical second derivatives, the Hessian $\tilde{H}$ must be reduced in size with the above projection. Specifically, the functional form of the constraints is

$$x'' = x''(x_{QMH}, x_{MMH}) = x_{QMH} + \alpha(x_{QMH} - x_{MMH}) \tag{30}$$

from which one can readily derive the relevant quantities for the projection in eqs 28−29. For instance, the scaling factor α is chosen to be 0.7261 for a link atom replacing a covalent single $C(sp^3)-C(sp^3)$ bond, it being the ratio of the equilibrium C−H and C−C distances in the CHARMM force field.[47] Note that the gradient correction in the last term of eq 29 drops out because of the linear relationship between x'' and x, i.e., $\partial^2 x''/\partial x \partial y = 0$ for x, y in $\{x_{QMH}, x_{MMH}\}$. It is clear that this projection affects only the Hessian elements of the host atoms and the link atom itself but no other Hessian elements. The projection can be performed within the QM code, since the derivatives of the classical parts vanish for these specific Hessian elements (e.g., $\tilde{H}^{x''y}_{class} = 0$).

As a last point, we mention a potential pitfall concerning the definition of the total energy of the QM/MM system with link atoms. The degree to which a link atom contributes to the total energy of the system is reason for discussion (see, e.g., refs 17, 48, and 49). As for normal-mode analysis, it is essential that both gradients and second derivatives conform to the chosen definition of the potential energy surface, that is, with the same constraints as in the energy minimization. This is readily satisfied in the Q-Chem/CHARMM interface. The user can specify the degree to which the link atom contributes to the total energy, and all derivatives are calculated accordingly. The subsequent elimination of the link atom coordinates, as in eqs 28−29, does not depend on the specific definition of the total energy of the system.

II.F. Long-Range Electrostatics. A common strategy, implemented in e.g. CP2K,[50] to reduce the computational cost of QM/MM electrostatic calculations is the introduction of a cutoff distance r_c. The Coulomb interaction between MM atoms and the QM region is neglected if the distance in between exceeds r_c. As a consequence, the number of perturbations decreases in the CPSCF equations of a QM/MM calculation, because a displacement of an MM atom which is too far away from the QM region will not affect the electronic cloud of the QM region. The full QM/MM Hessian still has its full $3N_{AT} \times 3N_{AT}$ dimension, but the reduced number of CPSCF equations facilitates its computation. In addition, the Hessian becomes more sparse, since Hessian elements $H_{A,B}$ between an MM atom A and a QM atom B are zero if they are beyond the cutoff.

Regardless of the tempting computational advantages, a cutoff strategy potentially introduces serious errors in the description of the molecular system. Electrostatics are long-range interactions with a $1/r_{AB}$ decay, leading to a $1/r_{AB}^3$ decay of the Hessian elements. A cutoff for the electrostatics leads to a shift of the potential energy and to modified vibrational frequencies with respect to the fully interacting system. It is to be anticipated that a cutoff distance r_c introduces errors on the order of $1/r_c^3$ in the Hessian elements.

The introduction of mobile blocks into the vibrational analysis has the advantage that the long-range electrostatics are not influenced. While the motion within a block is constrained, the description of the interatomic interactions is *in se* not altered. This is a major strength of the MBH approach with respect to coarse-graining methods: the MBH approach entails a correct description of long-range electrostatics and makes frequency calculations of large QM/MM systems feasible without invoking a cutoff technique.

III. Illustration of Computational Efficiency

The parallel implementation of MBH provides an efficient way to calculate vibrational frequencies of QM/MM systems. The reduced number of CPSCF equations results in computational profit on the level of both memory requirements and CPU time.

Formulas for the memory estimates of each step of sections II.B and II.D are included in Table 1 for a restricted closed-shell calculation, where n_o (n_v) denotes the number of occupied (virtual) orbitals and n_b denotes the number of basis functions. For open-shell calculations, the memory requirements double. In order to avoid load balance problems, the current implementation first solves the CPSCF equations for the QM atomic displacements followed by the MM atomic displacements.[51] Therefore, the peak memory requirement is $6N_{MM}n_b^2$ (reasonably assuming a larger number of MM atoms than QM atoms), whereas the needed disk storage still scales with $6N_{AT}n_b^2$. Parallellization over N_P processors reduces the memory by roughly a factor of N_P. The right column of Table 1 estimates the required memory per processor (besides the static work memory) for the calculation of the mobile block derivatives with N_P processors. The

Table 1. Memory Requirements Per Processor Expressed As the Number of Double Precision Floating-Point Numbers (1 number = 8 bytes) That Needs to Be Stored for Restricted Closed-Shell Hessian Calculations[a]

	full QM/MM Hessian	MB QM/MM Hessian
step 1	$3N_{MM} \times n_o n_v/N_P$	$d_{MM} \times n_o n_v/N_P$
step 2	$\ll 2 \times 3N_{MM} \times n_b^2/N_P$	$\ll 2 \times d_{MM} \times n_b^2/N_P$
step 3	$\propto n_b^2$	$\propto n_b^2$
step 4	$> 2 \times 3N_{MM} \times n_b^2/N_P$	$> 2 \times d_{MM} \times n_b^2/N_P$
step 5	$\propto n_b^2$	$\propto n_b^2$
peak	$2 \times 3N_{MM} \times n_b^2/N_P$	$2 \times d_{MM} \times n_b^2/N_P$

[a] N_P is the number of processors, n_b the number of basis functions, and n_o (n_v) the number of occupied (virtual) orbitals. The symbol $\ll$ means that the actual number is, in practice, much lower than the theoretical estimate. $\propto$ indicates the scaling, where the prefactor is independent of the system size. $>$ means that the actual number is higher than the theoretical estimate due to overhead such as the book keeping of variables. In the case of MBH, d_{MM} is the number of MM perturbations: $d_{MM} = 6N_{BL} + 3N_{MM,\,free}$, while the total number of perturbations is $d = d_{MM} + 3N_{AT}$.

peak memory is now equal to $2d_{MM}n_b^2/N_P$, where $d_{MM} = 6N_{BL} + 3N_{MM,\,free}$ is the number of MM perturbations.

To illustrate the efficiency, a small part (330 atoms) of the chorismate mutase enzyme (1COM in the Protein Data Bank) is taken as a test example.[52] At the B3LYP/6-31+G* level of theory, the number of occupied orbitals is $n_o = 59$, and the number of virtual orbitals is $n_v = 261$, totalling $n_b = 320$ basis functions. The full QM calculation would need roughly 1.5 TB of memory, which is not feasible at present. Now consider a QM/MM calculation where the system is divided into a 24 QM atoms and 306 MM atoms. The same level of theory, B3LYP/6-31+G*, is used for QM atoms while the MM atoms are described by the PARAM27 force field of CHARMM.[53,54] This QM/MM Hessian calculation only requires about 2.5 GB, a drastic reduction with respect to the full QM calculation.

In the next step, the system is divided into mobile blocks. By dividing the 306 MM atoms into 20 blocks, with one residue per block, a mere 0.5 GB is sufficient for the MBH calculation on one processor. This illustrates that MBH allows for vibrational analysis on larger systems than is feasible with the full Hessian. For example, the largest system in which the full Hessian calculation fits on a single processor with a typical 8 GB of memory consists of roughly 24 QM atoms and 600 MM partial charges. The largest system for the approximate MBH calculation has 24 QM atoms and 4500 MM atoms, assuming an average of 15 atoms per block.

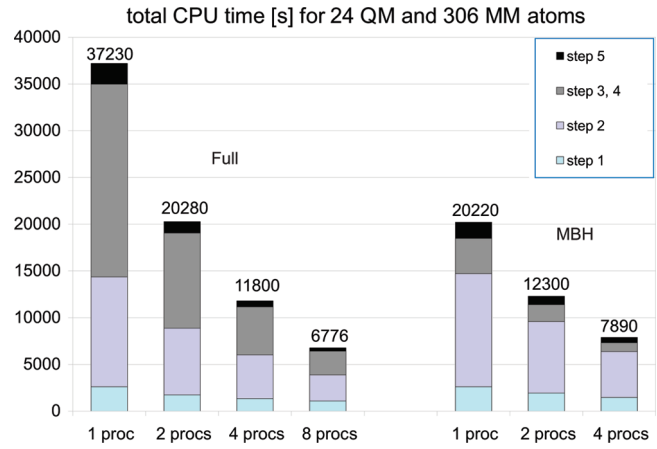


Figure 3. Chorismate mutase test system: CPU times for a QM/MM system with 24 QM atoms and 306 MM atoms. Parallellization speeds up the QM/MM calculatation: the full Hessian is calculated on 1, 2, 4, and 8 processors (left); the mobile block Hessian on 1, 2, and 4 processors (right).

Thus, MBH increases the size of the QM/MM systems that can be addressed by 7.5 times. Even larger system sizes are feasible when grouping multiple residues per block. In addition, parallellization of the code further reduces the required memory per processor.

Figure 3 shows the CPU times of the frequency calculation for the chorismate mutase test system, described with the same QM/MM description of 24 QM atoms and 306 MM atoms. The timings of the full Hessian calculation are compared to those of the MBH, where again the 306 MM atoms are divided into 20 blocks. The MBH CPU time on one processor (20 220 s) reduces to 54% of the full Hessian CPU time on one processor (37 230 s). The CPU times of steps 3 and 4 are affected the most, while the CPU time of step 2 remains unaltered or might even increase. The latter is due to the use of an iterative subspace algorithm for solving the CPSCF equations. The number of trial vectors in the CPSCF solution subspace is largely unaffected by blocking the MM atoms. In our particular example, the number of basis vectors is actually slightly increased because of the mixing of different atomic displacements in the blocks. This leads to a subsequent small increase in the execution time for the CPSCF step, which is almost strictly proportional to the number of basis vectors. For the full Hessian calculation, a speedup of a factor of 5.5 is realized when using eight processors instead of one processor. For the MBH calculation, the speedup is 2.6 when using four processors instead of one processor. This shows that MBH indeed reduces the computational efficiency, and the parallel implementation even more. Of course, the effective speedup depends on the block choice, where larger blocks lead to more impressive speedups.

IV. Application: The Bortezomib Drug

IV.A. Oxidative Deboronation of Bortezomib. Boronic acids (R-B(OH)$_2$) play an important role in a variety of medical applications due to the ability of boron to mimic the tetrahedral transition state of an sp^3 hybridized carbon. A particularly interesting example is the drug bortezomib (originally codenamed PS-341, marketed as Velcade),[29,30] which is

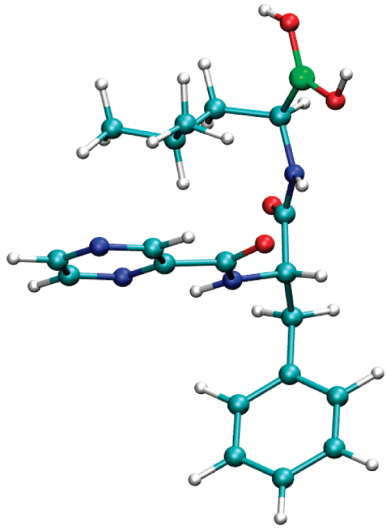


Figure 4. The drug bortezomib: 3D model of the reactant (REA).

used to treat multiple myeloma and mantle cell lymphoma, two types of hematologic cancer. Bortezomib is a boronic acid analog of a Phe−Leu dipeptide coupled to a 2-carboxyl-pyrazine group (Figure 4) that binds to the catalytic site of the 26S proteasome with high specificity.[55,56] The proteasome, a large multicatalyic protease complex, regulates protein expression and degradation of ubiquitinylated proteins, cleaning abnormal or misfolded proteins from the cell. Inhibition of this cellular pathway by bortezomib ultimately results in apoptosis due to an accumulation of damaged or misfolded proteins in the cell through a number of possible mechanisms.[57]

The chemical activity of bortezomib is largely due to the boronic acid moiety, which appears to bind with the active site N-terminal threonine residue of the proteasome.[58−60] In a recent article, Larkin et al. reported the results of a computational study of the model system boroglycine ($H_2N-CH_2-B(OH)_2$), using H_2O_2 and H_2O as reactive oxygen species.[61] The oxidative deboronation, which is suggested as the principal pathway for the metabolism of bortezomib, is found to be exothermic and endothermic for the reactions with H_2O_2 and H_2O, respectively. With the computational improvements in Q-Chem/CHARMM, we can now systematically study the full bortezomib molecule (53 atoms) instead of the smaller model system (11 atoms).

Figure 5 illustrates the oxidative deboronation reaction, where an oxygen cleaves the boron acid and takes the position of the boron. The vibrational free energy difference ΔG_{vib} of the reaction is studied since this quantity is calculated with the vibrational frequencies. Two reactive oxygen reagents are considered: H_2O_2 and H_3C-OH. Methanol is chosen as the second reagent because it better describes—compared to water in ref 61— the alcohol group of the threonine residue of the 26S proteasome on which the bortezomib molecule is believed to bind. The respective products are an alcohol and an ether:

$$R-B-(OH)_2 + H_2O_2 \rightarrow R-OH + B-(OH)_3$$

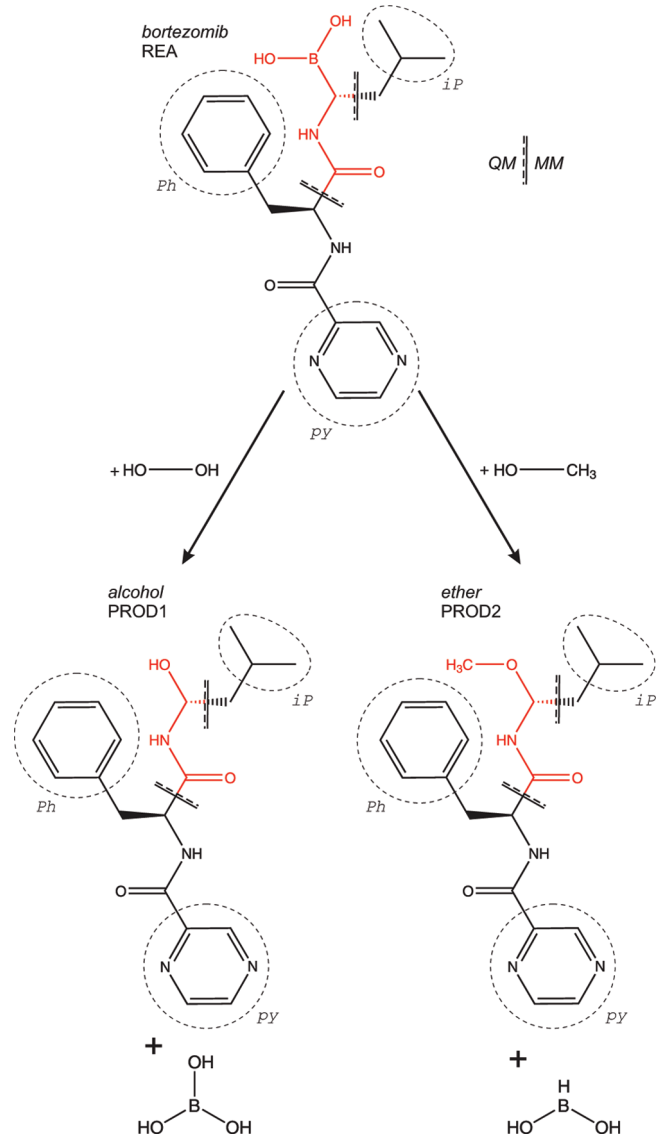


Figure 5. Bortezomib—Structure of the drug bortezomib with indication of the QM and MM regions. QM atoms are colored red. Link atoms are placed at the QM/MM border. Oxidative deboronation of bortezomib (REA) with the oxygen reagens H_2O_2 or CH_3-OH results in an alcohol (PROD1) or ether (PROD2) product, respectively. The mobile blocks Ph, py, and iP are indicated with a dashed line.

$$R-B-(OH)_2 + H_3C-OH \rightarrow R-O-CH_3 + H-B-(OH)_2$$

where R designates the Phe−Leu moieties coupled to the pyrazine. The bortezomib reactant will be referred to as REA and the alcohol and ether product as PROD1 and PROD2, respectively.

With this application, we aim at presenting a workable model for the computation of vibrational frequencies in a QM/MM approach with the inclusion of mobile blocks in the MM region. We restrict ourselves to the thermochemistry of the reactants versus the products, as the reaction path leads along several intermediate transition states.[61] A profound assessment can only be done if we dispose of a high-level benchmark study involving accurate optimized geometries for the bortezomib structure. Therefore, section IV.B is devoted to the creation of the QM and QM/MM geometries.

Table 2. Bortezomib−RMSD in Å between QM Geometries Optimized at Different Levels of Theory[a]

	QM geometries	B3LYP	PBE	B3LYP-D	PBE-D
REA	B3LYP	−			
	PBE	0.10	−		
	B3LYP-D	0.57	0.50	−	
	PBE-D	0.68	0.60	0.12	−
	RI-MP2/cc-pvtz	0.58	0.51	0.09	0.12
PROD1	B3LYP	−			
	PBE	0.64	−		
	B3LYP-D	0.75	0.34	−	
	PBE-D	0.82	0.42	0.11	−
	RI-MP2/cc-pvtz	0.78	0.45	0.14	0.09
PROD2	B3LYP	−			
	PBE	0.07	−		
	B3LYP-D	0.91	0.86	−	
	PBE-D	0.97	0.93	0.09	−
	RI-MP2/cc-pvtz	0.94	0.89	0.16	0.14

[a] The basis set is 6-311++G(d,p) except for RI-MP2. REA refers to the reactant bortezomib, while PROD1 and PROD2 refer to the alcohol and ether products formed after oxidative deboronation, as indicated in Figure 5.

Section IV.C focuses on the calculation of the QM and QM/MM frequencies, either derived from full Hessians or from MBH. Section IV.D compares the QM/MM difference in the vibrational Gibbs free energy with the benchmark full QM value. Moreover, we discuss the influence of the introduction of mobile blocks in the MM region.

IV.B. Creating QM and QM/MM Geometries. The benchmark geometries are generated by performing a full QM calculation with Q-Chem. The input geometry is taken from the Protein Data Bank (2F16 in the Protein Data Bank) where the proteasome 26S and water are removed to select the relevant conformation. The geometry of each structure is first optimized in order to calculate frequencies subsequently using analytical second derivatives. Since it is not required that the gradient within a given block be zero in order to apply MBH, the geometry optimization could also be done with rigid blocks, where the rigidity is imposed via the SHAPES facility in CHARMM. To increase the accuracy, the convergence criteria and numerical accuracy options are set slightly tighter than is typical for a plain geometry optimization: the tolerance for the gradient is reduced to 0.00015 hartree/Bohr, and the tighter (75 302) atomic integration grid is chosen with 75 radial points and 302 Lebedev angular points for each atom. The B3LYP[62,63] and PBE (=PBE1PBE)[64−66] levels of theory are used for this example with the 6-311++G(d,p) basis set, as well as the B3LYP-D and PBE-D functionals where an empirical correction term is added to account for dispersion effects.[67] In addition, the structure is optimized at the more expensive RI-MP2/cc-pVTZ[68,69] level to validate the accuracy of the B3LYP(-D) and PBE(-D) geometries. A frequency calculation is however currently not practical at the RI-MP2 level of theory because its second derivatives implementation is based on the numerical differentiation of the analytical gradient. Table 2 lists the mass-weighted root-mean-square distances (RMSD) between the QM geometries optimized at different levels of theory, which are calculated on the basis of the non-hydrogen atoms after aligning the structures. Structures with an RMSD below 0.25 Å are considered to have close-lying geometries; when the RMSD is higher, the structures are considered less similar. The RMSD between B3LYP and PBE structures is low, except in the isolated case of the PROD1 product. Similarly, the B3LYP-D and PBE-D geometries lie close to each other, but they differ from the geometries without dispersion. In all cases, the B3LYP-D and PBE-D geometries lie closer to the RI-MP2 geometries than B3LYP and PBE, and preference should be given to DFT methods including dispersion. Visualization of the structures shows that the distance between the Leu and Phe moieties decreases under the influence of the dispersion forces. Apparently, the dispersion in this rather large molecular system plays a more important role on the geometry than the level of theory.

For the QM/MM calculation, the REA, PROD1, and PROD2 molecules are divided into a QM and MM region, whereas the small H_2O_2 and CH_3-OH reactants and the $B-(OH)_3$ and $H-B-(OH)_2$ products are still described completely on the QM level. The MM region consists of 42 atoms: the phenyl group (Phe), the iso-butyl group (Leu), the pyrazine (ring with nitrogens), and some neighboring atoms, as shown in Figure 5. The reactive site is made part of the QM region, colored red in Figure 5. In addition, the amide bond is chosen to belong entirely to the QM region, since preliminary tests in which the QM/MM border crosses the amide bond turned out to break the partially delocalized nature of the amide bond with even a nonplanar geometry in some cases. This brings the number of QM atoms to 11, 8, and 11 for the REA, PROD1, and PROD2 molecules, respectively. Each structure has two link atoms where the QM/MM border cuts through covalent C−C bonds, as indicated in Figure 5. The combination of the LONEPAIR and SHAKE commands[70] of CHARMM keeps each link atom colinear with its MM and QM host at a relative distance of 0.7261 (see eq 30) during the geometry optimization.

In this QM/MM calculation, the same QM functionals are used as in the full QM case (the B3LYP, PBE, B3LYP-D, or PBE-D level with the 6-311++G(d,p) basis set), while the MM force field is based on the PARAM27 parameter set[53,54] of CHARMM. The root-mean-square distances (RMSD) between the QM/MM geometries optimized at these four levels of theory are summarized in Table 3. As in the QM calculations, the B3LYP and PBE structures are similar, and the B3LYP-D and PBE-D structures also lie close to each other. Contrary to the full QM calculations, the comparison of B3LYP and PBE with B3LYP-D and PBE-D shows that the inclusion of dispersion interactions only has a minor influence on the geometry. The reason is that the dispersion contribution is limited to the small subset of QM atoms; hence it barely affects the relative orientation between the Leu and Phe moieties. This behavior is confirmed by comparing the QM/MM structure with its respective QM structure, of which the RMSD is also included in Table 3. Indeed, QM/MM structures calculated in the absence of dispersion lie closer to their QM counterpart than structures including dispersion. For instance, the RMSD between the

Table 3. Bortezomib−RMSD in Å between QM/MM Geometries Optimized at Different Levels of Theory[a]

	QM/MM geometries	B3LYP	PBE	B3LYP-D	PBE-D
REA	B3LYP	–			
	PBE	0.05	–		
	B3LYP-D	0.11	0.07	–	
	PBE-D	0.15	0.11	0.05	–
	correspondingQM	0.18	0.20	0.43	0.50
PROD1	B3LYP	–			
	PBE	0.14	–		
	B3LYP-D	0.16	0.03	–	
	PBE-D	0.23	0.09	0.07	–
	correspondingQM	0.18	0.60	0.63	0.62
PROD2	B3LYP	–			
	PBE	0.16	–		
	B3LYP-D	0.18	0.04	–	
	PBE-D	0.12	0.05	0.07	–
	correspondingQM	0.24	0.16	0.87	0.88

[a] In addition, each QM/MM geometry is compared with its respective QM geometry calculated at the same level of theory. REA refers to the reactant bortezomib, while PROD1 and PROD2 refer to the alcohol and ether products formed after oxidative deboronation, as indicated in Figure 5.

QM/MM and QM structure is 0.18 Å with the B3LYP functional, while it is 0.43 Å with the B3LYP-D functional.

IV.C. Frequency Calculations. Hessians and gradients are calculated analytically at the same level of theory as the geometry optimization. The Hessian is diagonalized after mass-weighting to obtain the frequencies and vibrational modes. In this paper, we used simultaneously the program TAMkin[71] to derive frequencies, modes, and thermodynamic properties. With its batch processing features, it provides a handy interface to extract the relevant molecular information from the large number of Q-Chem/CHARMM output files, to compute the frequencies, to write mode trajectory files for visualization of the vibrational eigenmodes, and to derive the Gibbs free energy differences.

Frequencies are derived from three different types of Hessians:

A. QM Full Hessian. The frequency run (and geometry optimization) is performed with a QM description. The full $3N_{AT} \times 3N_{AT}$ Hessian is calculated with Q-Chem.

B. QM/MM Full Hessian. The frequency run (and geometry optimization) is performed with a QM/MM description. The full $3N_{AT} \times 3N_{AT}$ Hessian is calculated with Q-Chem/CHARMM. CHARMM can derive MBH frequencies from this full Hessian.

C. QM/MM Mobile Block Hessian, Reduced CPSCF. The frequency run (and geometry optimization) is performed with a QM/MM description. The mobile block Hessian of reduced dimension $d \times d$ is calculated by Q-Chem/CHARMM with the reduced number of CPSCF equations and diagonalized in CHARMM, directly leading to the MBH frequencies.

In cases A and B, a Hessian of full size $3N_{AT} \times 3N_{AT}$ is constructed. Its diagonalization is performed by Q-Chem (case A), by CHARMM (case B), or by TAMkin (case A or B) and will be referred to as the full Hessian vibrational analysis (FHVA) in the remainder of the discussion. From these full Hessians, one can also derive MBH frequencies through projection and a gradient correction, as explained in ref 26 (cases A and B) and implemented in CHARMM[19] and in TAMkin.[71] But the direct method to attain MBH frequencies is case C, where frequencies evolve directly from the diagonalization of the mobile block Hessian itself, without prior construction of the full Hessian. This is the new implementation which is the subject of this paper and which is now available via Q-Chem/CHARMM.

The calculations are performed on four processors with the parallel version of Q-Chem and Q-Chem/CHARMM. The QM Hessian computation in case A is the most time-consuming one (~3 days). The QM/MM Hessian calculations in cases B and C are considerably faster, taking approximately 0.5% of the time for a full QM calculation (~30 min). The speedup realized by the reduced CPSCF implementation versus a full CPSCF QM/MM implementation is moderate since the number of MM atoms and blocks is rather small in the system under study compared to the chorismate mutase test system of section III.

In the current implementation, the mobile blocks should be part of the MM region and should not contain any MM host atom of the link atoms. Figure 5 proposes three plausible blocks, of which the internal motions are suspected not to matter when estimating the vibrational free energy difference: the phenyl group (Ph), the iso-propyl group (iP), and the pyrazine (py). Frequencies are calculated with one block or with multiple blocks simultaneously. For instance, the method MBH_{Ph} indicates that the vibrational analysis is performed assuming that the atoms of block Ph vibrate coherently as a whole. In the method labeled $MBH_{Ph, iP, py}$, the system contains three blocks (Ph, iP, py), while the remaining atoms can vibrate individually. The introduction of blocks reduces the number of frequencies; for instance, 87 frequencies remain for the bortezomib reactant with three blocks, which is 55% of the original 159 frequencies.

The accuracy of the vibrational frequencies is largely influenced by the geometry convergence criteria and the quality of the Hessian elements, which mainly depends on the numerical integration grid for calculating the electron integrals in the CPSCF equations. A good assessment of the accuracy is the value of the lowest six eigenvalues of the Cartesian Hessian. In principle, those should be zero at a minimum or maximum energy point because of the invariance of the potential energy surface of a gas phase molecule under global translations and rotations.[26,72,73] Global rotations with a nonzero frequency are caused by a poor geometry convergence, such that the Eckart conditions are not fulfilled at the reference point and the global rotations may mix up with the internal vibrations. Global translations with a nonzero frequency signal inaccurate Hessian elements. The influence of the inaccuracies is validated by first projecting out the global translations/rotations from the full Hessian before diagonalization, thus creating six "hard" zero eigenfrequencies and removing any translational/rotational contribution in the low lying eigenmodes. This approach is equivalent to imposing Eckart constraints on the vibrational motions[1] and will therefore be denoted with the superscript "Eck". The Eckart projection is performed by TAMkin or by the RAISE option in CHARMM. Each standard method, FHVA or MBH, thus has a corresponding method $FHVA^{Eck}$

or MBH^{Eck} where the six lowest frequencies are hard zeros. The Eckart projection mainly affects the low frequency spectrum. The difference between the standard method and these Eckart methods is a measure of the Hessian's accuracy.

IV.D. Discussion. A first point of interest is the influence of MBH on the individual frequencies and modes. Such detailed studies have been performed recently in, e.g., refs 25, 27, 37, 74, and 75. It was found that the block choice determines which local modes and/or global modes are well described by MBH. For the bortezomib system under study, all atoms of the reactive site are considered completely mobile free atoms. The spectator groups are fairly rigid during the reaction, and their internal geometry can be kept fixed. With this plausible block choice, a reasonable similarity is to be expected between MBH modes and frequencies and the benchmark full Hessian results. This can be verified by calculating the overlap between the modes, a number lying between 0 and 100%, defined as

$$O_{ij} = |\langle \nu_i^{MBH} | \nu_j \rangle|^2 \tag{31}$$

where $|\nu_i^{MBH}\rangle$ is the ith mass-weighted MBH mode with frequency ν_i^{MBH} and $|\nu_j\rangle$ is the jth mass-weighted FHVA mode with frequency ν_j. The overlap data of REA are plotted in Figure 6, where the QM/MM MBH is derived from the full QM/MM Hessian by projection (case B of ssection IV.C). A dark dot indicates a high overlap between the modes, and a dot located close to the diagonal of the plot indicates that the corresponding frequencies are almost equal. The plot shows good agreement between the FHVA modes of REA with the MBH^{Ph} and $MBH^{Ph,\ iP,\ py}$ modes, and an excellent agreement between FHVA and MBH for frequencies below 250 cm^{-1}.

A second point of interest is the thermodynamic quantities derived from the frequencies. The frequencies serve directly as input quantities for the vibrational free energy G_{vib}, which makes it an interesting parameter for studying the influence of the MBH model. The MBH however reduces the number of frequencies; hence a better quantity is the *difference* in vibrational free energy ΔG_{vib} between the products and the reactants, calculated as

$$\Delta G_{vib} = G_{vib}(PROD1) + G_{vib}(B(OH)_3) - G_{vib}(REA) - G_{vib}(H_2O_2) \tag{32}$$

$$\Delta G_{vib} = G_{vib}(PROD2) + G_{vib}(HB(OH)_2) - G_{vib}(REA) - G_{vib}(H_3COH) \tag{33}$$

In the harmonic oscillator approximation, the vibrational free energy G_{vib} is derived from the vibrational partition function Q_{vib} by the well-known relation $G_{vib} = -k_B T \ln Q_{vib}$, where k_B is the Boltzmann constant and T the temperature. The vibrational partition function is built from the individual contributions of the harmonic frequencies.[76−79] The entropic part of the vibrational free energy, $-T\Delta S_{vib}$, is also reported, which is derived from the relation $S_{vib} = k_B(\partial \ln Q_{vib})/(\partial T)$. When the harmonic oscillators are treated classically (high temperature limit) instead of quantum mechanically, the free energy difference is purely entropic.

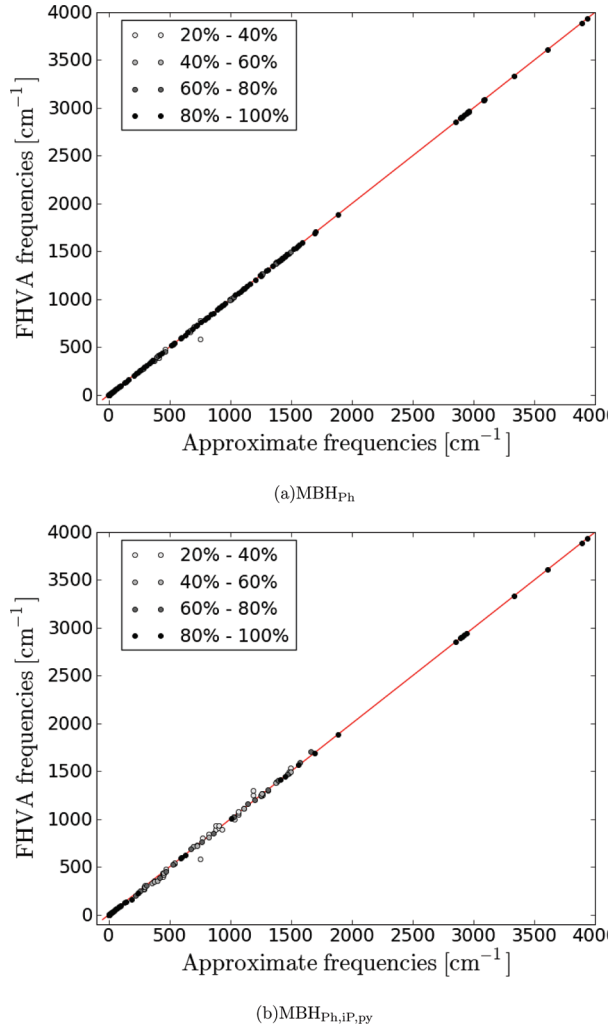


Figure 6. Bortezomib—Evaluation of the MBH frequencies/modes of REA on the basis of the QM/MM Hessian (case B). The overlap $O_{ij} = |\langle \nu_i^{MBH} | \nu_j \rangle|^2$ between the mass-weighted MBH modes $|\nu_i^{MBH}\rangle$ and FHVA modes $|\nu_j\rangle$ is plotted as a function of the respective frequencies. A dark colored dot indicates a high overlap between the modes; overlap values below 20% are not shown.

The alternative would be to derive the free energy differences from a molecular dynamics (MD) simulation, which performs a more realistic sampling of the reaction coordinate. In this paper, only the harmonic limit is considered, which amounts to sampling a local harmonic approximation of the potential energy surface. Even in cases where the harmonic limit is insufficient to accurately describe the reaction process, one can still learn from the harmonic limit result by comparing it to explicit MD sampling. Since MD basically solves Newton's classical equations of motion and not the time-dependent Schrodinger equation, MD results should be compared with the results derived from the classical oscillators instead of the quantum oscillators partition function. For completeness, we have therefore calculated some classical oscillator results besides the quantum oscillator results.

Tables 4−7 list the vibrational free energies and the vibrational entropic contributions, calculated with quantum/classical oscillators, and without/with Eckart projection. The

Table 4. Deboronation of Bortezomib with the Oxygen Reagents H_2O_2[a]

		B3LYP				PBE			
Hessian	NMA method	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$
case A	FHVA	2.40	2.47	1.65	1.88	2.50	2.24	1.57	1.77
QM	MBH_{Ph}	2.40	2.47	1.65	1.88	2.48	2.23	1.57	1.77
full	MBH_{iP}	2.33	2.40	1.68	1.88	2.30	2.01	1.41	1.59
	MBH_{py}	2.40	2.47	1.66	1.88	2.50	2.23	1.56	1.77
	$MBH_{Ph,\,iP}$	2.33	2.40	1.68	1.88	2.28	2.00	1.41	1.58
	$MBH_{Ph,\,py}$	2.40	2.47	1.65	1.88	2.48	2.22	1.56	1.76
	$MBH_{iP,\,py}$	2.33	2.40	1.68	1.89	2.29	2.00	1.40	1.58
	$MBH_{Ph,\,iP,\,py}$	2.33	2.40	1.68	1.88	2.27	1.99	1.40	1.57
case B	FHVA	2.77	2.81	2.16	2.33	2.93	2.96	2.28	2.47
QM/MM	MBH_{Ph}	2.77	2.81	2.16	2.33	2.94	2.97	2.28	2.47
full	MBH_{iP}	2.76	2.80	2.16	2.33	2.95	2.98	2.32	2.50
	MBH_{py}	2.78	2.82	2.16	2.34	2.92	2.95	2.27	2.46
	$MBH_{Ph,\,iP}$	2.76	2.80	2.16	2.33	2.96	2.99	2.32	2.51
	$MBH_{Ph,\,py}$	2.78	2.82	2.16	2.34	2.93	2.96	2.28	2.47
	$MBH_{iP,\,py}$	2.76	2.81	2.17	2.34	2.94	2.97	2.31	2.49
	$MBH_{Ph,\,iP,\,py}$	2.76	2.81	2.17	2.34	2.95	2.98	2.32	2.50
case C	MBH_{Ph}	2.81	2.87	2.21	2.38	2.97	3.03	2.34	2.53
QM/MM	MBH_{iP}	2.82	2.87	2.23	2.40	2.99	3.03	2.36	2.55
mobile block	MBH_{py}	2.79	2.84	2.18	2.35	2.93	2.97	2.28	2.47
(several)	$MBH_{Ph,\,iP}$	2.79	2.83	2.19	2.36	2.97	3.00	2.34	2.52
	$MBH_{Ph,\,py}$	2.81	2.84	2.18	2.36	2.94	2.97	2.29	2.48
	$MBH_{iP,\,py}$	2.80	2.84	2.20	2.37	2.98	3.01	2.34	2.53
	$MBH_{Ph,\,iP,\,py}$	2.80	2.83	2.19	2.36	2.99	3.02	2.35	2.54

[a] The vibrational free energy difference of eq 32 and its entropic part are calculated (in kcal/mol) with several NMA models. The superscript "cl" indicates the use of classical instead of quantum oscillators. The superscript "Eck" indicates that Eckart conditions are applied. MBH and MBH^{Eck} frequencies are derived from the QM full Hessian (case A) and the QM/MM full Hessian (full CPSCF, case B) with CHARMM or TAMkin or are obtained from the direct QM/MM mobile block Hessian (reduced CPSCF, a different Hessian for each block choice, case C). The basis set is 6-311++G(d,p).

Table 5. Deboronation of Bortezomib with the Oxygen Reagens H_2O_2 Continued[a]

		B3LYP-D				PBE-D			
Hessian	NMA method	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$
case A	FHVA	1.78	1.87	1.12	1.33	2.67	2.36	1.50	1.74
QM	MBH_{Ph}	1.86	1.94	1.13	1.36	2.69	2.39	1.52	1.76
full	MBH_{iP}	1.99	2.08	1.24	1.46	2.77	2.73	1.85	2.08
	MBH_{py}	1.80	1.89	1.13	1.34	2.62	2.31	1.49	1.72
	$MBH_{Ph,\,iP}$	2.07	2.15	1.25	1.49	2.80	2.75	1.87	2.10
	$MBH_{Ph,\,py}$	1.88	1.96	1.13	1.37	2.64	2.33	1.51	1.74
	$MBH_{iP,\,py}$	2.01	2.10	1.25	1.48	2.72	2.67	1.84	2.06
	$MBH_{Ph,\,iP,\,py}$	2.09	2.17	1.26	1.50	2.75	2.70	1.86	2.08
case B	FHVA	2.75	2.76	1.99	2.20	2.28	2.26	1.49	1.71
QM/MM	MBH_{Ph}	2.69	2.70	1.93	2.14	2.28	2.26	1.50	1.71
full	MBH_{iP}	2.68	2.69	1.94	2.15	2.35	2.36	1.61	1.82
	MBH_{py}	2.74	2.75	1.98	2.19	2.28	2.26	1.49	1.71
	$MBH_{Ph,\,iP}$	2.62	2.63	1.88	2.08	2.35	2.37	1.62	1.83
	$MBH_{Ph,\,py}$	2.69	2.69	1.92	2.13	2.29	2.26	1.50	1.71
	$MBH_{iP,\,py}$	2.67	2.68	1.93	2.14	2.35	2.36	1.61	1.82
	$MBH_{Ph,\,iP,\,py}$	2.61	2.62	1.87	2.07	2.35	2.36	1.61	1.82
case C	MBH_{Ph}	2.73	2.75	1.97	2.18	2.29	2.28	1.51	1.72
QM/MM	MBH_{iP}	2.71	2.72	1.96	2.17	2.41	2.41	1.65	1.86
mobile block	MBH_{py}	2.76	2.77	1.99	2.21	2.32	2.28	1.52	1.73
(several)	$MBH_{Ph,\,iP}$	2.65	2.65	1.90	2.11	2.38	2.39	1.64	1.85
	$MBH_{Ph,\,py}$	2.70	2.71	1.94	2.15	2.30	2.25	1.49	1.70
	$MBH_{iP,\,py}$	2.69	2.70	1.95	2.16	2.36	2.36	1.61	1.81
	$MBH_{Ph,\,iP,\,py}$	2.65	2.66	1.90	2.11	2.38	2.39	1.64	1.85

[a] See caption of Table 4.

three parts in each table correspond to the three different cases of Hessians as explained above. Variations $\delta\Delta G_{vib}$ in a particular column in the table are due to the description level (QM versus QM/MM) and the NMA models (FHVA versus MBH). Table 8 summarizes these variations $\delta\Delta G$ as well as deviations caused by the functional and the application of the Eckart conditions. The deviations will now be discussed in detail on the basis of the results obtained from quantum oscillators. Classical oscillator results, indicated with a superscript "cl" in Tables 4−7, lead to conclusions

Table 6. Deboronation of Bortezomib with the Oxygen Reagens CH_3OH[a]

		B3LYP				PBE			
Hessian	NMA method	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$
case A	FHVA	−0.13	−0.03	0.62	0.83	−1.11	−1.15	−0.41	−0.20
QM	MBH_{Ph}	−0.12	−0.03	0.62	0.83	−1.14	−1.19	−0.43	−0.23
full	MBH_{iP}	−0.23	−0.13	0.63	0.81	−1.33	−1.34	−0.44	−0.28
	MBH_{py}	−0.14	−0.04	0.62	0.83	−1.12	−1.16	−0.41	−0.20
	$MBH_{Ph,\,iP}$	−0.22	−0.12	0.63	0.81	−1.36	−1.37	−0.46	−0.31
	$MBH_{Ph,\,py}$	−0.13	−0.04	0.62	0.84	−1.15	−1.19	−0.44	−0.23
	$MBH_{iP,\,py}$	−0.23	−0.14	0.63	0.81	−1.33	−1.35	−0.44	−0.28
	$MBH_{Ph,\,iP,\,py}$	−0.23	−0.13	0.63	0.82	−1.36	−1.38	−0.47	−0.31
case B	FHVA	0.38	0.43	1.26	1.42	0.16	0.20	1.12	1.26
QM/MM	MBH_{Ph}	0.33	0.38	1.21	1.36	0.16	0.19	1.11	1.25
full	MBH_{iP}	0.43	0.47	1.33	1.47	0.18	0.21	1.14	1.28
	MBH_{py}	0.39	0.43	1.27	1.42	0.17	0.21	1.13	1.26
	$MBH_{Ph,\,iP}$	0.38	0.43	1.28	1.42	0.17	0.20	1.14	1.27
	$MBH_{Ph,\,py}$	0.34	0.38	1.21	1.37	0.17	0.20	1.12	1.26
	$MBH_{iP,\,py}$	0.43	0.47	1.33	1.47	0.18	0.22	1.15	1.29
	$MBH_{Ph,\,iP,\,py}$	0.38	0.43	1.28	1.43	0.18	0.21	1.14	1.28
case C	MBH_{Ph}	0.37	0.44	1.27	1.42	0.20	0.27	1.18	1.32
QM/MM	MBH_{iP}	0.48	0.53	1.37	1.52	0.21	0.25	1.18	1.32
mobile block	MBH_{py}	0.40	0.44	1.27	1.42	0.18	0.21	1.13	1.27
(several)	$MBH_{Ph,\,iP}$	0.41	0.45	1.29	1.44	0.18	0.21	1.14	1.28
	$MBH_{Ph,\,py}$	0.35	0.39	1.23	1.38	0.18	0.22	1.14	1.28
	$MBH_{iP,\,py}$	0.46	0.50	1.35	1.50	0.22	0.26	1.19	1.32
	$MBH_{Ph,\,iP,\,py}$	0.43	0.47	1.32	1.47	0.22	0.25	1.18	1.32

[a] See caption of Table 4.

Table 7. Deboronation of Bortezomib with the Oxygen Reagens CH_3OH Continued[a]

		B3LYP-D				PBE-D			
Hessian	NMA method	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$	ΔG_{vib}	ΔG_{vib}^{Eck}	$-T\Delta S_{vib}^{Eck}$	$\Delta G_{vib}^{Eck,\,cl}$
case A	FHVA	0.63	0.65	1.30	1.51	−0.59	−0.56	0.36	0.50
QM	MBH_{Ph}	0.58	0.59	1.29	1.48	−0.59	−0.56	0.36	0.50
full	MBH_{iP}	0.75	0.77	1.42	1.64	−0.60	−0.57	0.36	0.50
	MBH_{py}	0.61	0.63	1.30	1.51	−0.62	−0.59	0.35	0.49
	$MBH_{Ph,\,iP}$	0.69	0.71	1.40	1.61	−0.61	−0.58	0.36	0.50
	$MBH_{Ph,\,py}$	0.55	0.57	1.29	1.47	−0.62	−0.59	0.36	0.49
	$MBH_{iP,\,py}$	0.72	0.74	1.42	1.63	−0.63	−0.60	0.35	0.49
	$MBH_{Ph,\,iP,\,py}$	0.66	0.68	1.40	1.60	−0.63	−0.60	0.35	0.48
case B	FHVA	0.59	0.59	1.27	1.47	0.37	0.38	1.17	1.35
QM/MM	MBH_{Ph}	0.65	0.66	1.33	1.53	0.37	0.38	1.17	1.35
full	MBH_{iP}	0.64	0.65	1.34	1.53	0.37	0.38	1.19	1.36
	MBH_{py}	0.59	0.60	1.29	1.48	0.37	0.38	1.18	1.35
	$MBH_{Ph,\,iP}$	0.70	0.71	1.40	1.59	0.37	0.38	1.18	1.36
	$MBH_{Ph,\,py}$	0.66	0.67	1.34	1.54	0.37	0.38	1.17	1.35
	$MBH_{iP,\,py}$	0.65	0.66	1.35	1.54	0.37	0.39	1.19	1.36
	$MBH_{Ph,\,iP,\,py}$	0.71	0.72	1.41	1.60	0.37	0.39	1.19	1.36
case C	MBH_{Ph}	0.68	0.70	1.37	1.57	0.42	0.44	1.22	1.41
QM/MM	MBH_{iP}	0.68	0.68	1.37	1.56	0.43	0.44	1.23	1.41
mobile block	MBH_{py}	0.61	0.62	1.30	1.49	0.39	0.40	1.18	1.36
(several)	$MBH_{Ph,\,iP}$	0.73	0.74	1.42	1.61	0.39	0.40	1.20	1.37
	$MBH_{Ph,\,py}$	0.67	0.68	1.36	1.55	0.39	0.41	1.20	1.38
	$MBH_{iP,\,py}$	0.67	0.68	1.37	1.56	0.40	0.42	1.22	1.39
	$MBH_{Ph,\,iP,\,py}$	0.75	0.76	1.45	1.64	0.41	0.42	1.22	1.39

[a] See caption of Table 4.

similar to those of the quantum oscillator results and are not discussed separatedly.

First, consider the FHVA results of the reaction of bortezomib with H_2O_2 to form the alcohol product PROD1 (Tables 4 and 5). The vibrational contribution to the reaction free energy is unfavorable, since the ΔG_{vib} values are positive. The QM ΔG_{vib} [FHVA] results range from 1.78 to 2.40 kcal/mol, and the QM/MM ΔG_{vib} [FHVA] results range from 2.28 to 2.93 kcal/mol. Both the internal energy and the entropic part are positive for all four functionals. The results of the reaction with methanol to form the ether product PROD2 are more complex (Tables 6 and 7). The QM ΔG_{vib} [FHVA] results range from −1.11 to 0.63 kcal/mol, where only the B3LYP-D functional has a positive value, whereas the QM/MM ΔG_{vib} [FHVA] results range from 0.16 to 0.59 kcal/mol, all being positive. The vibrational

Table 8. Bortezomib—Average Deviations of the Absolute Value of the Vibrational Free Energy (δG_{vib}) and Average Deviations of the Vibrational Free Energy Difference ($\delta \Delta G_{vib}$) Caused by Several Calculation Parameters: The QM versus QM/MM Description, The Choice of Functional (B3LYP, PBE, B3LYP-D, PBE-D), The NMA Model (FHVA, FHVA^{Eck}, MBH, MBH^{Eck}) and the Mobile Block Hessian Implementation (case B versus case C)

source of deviation	δG_{vib} [kcal/mol]	
FHVA vs FHVA^{Eck}	0.18 (QM)	0.02 (QM/MM)
source of deviation	**$\delta \Delta G_{vib}$ [kcal/mol]**	
QM vs QM/MM (FHVA)	0.62	
functional (FHVA)	0.53 (QM)	0.20 (QM/MM)
FHVA vs FHVA^{Eck}	0.11 (QM)	0.02 (QM/MM)
MBH vs MBH^{Eck}	0.10 (QM)	0.03 (QM/MM)
MBH vs FHVA	0.09 (QM)	0.04 (QM/MM)
MBH^{Eck} vs FHVA^{Eck}	0.11 (QM)	0.04 (QM/MM)
case B vs case C (MBH)	–	0.03 (QM/MM)

effect on the reaction kinetics in the QM description is therefore unclear for this reaction when comparing the four QM functionals. The internal energy difference of the quantum oscillators is always negative, but the entropic part depends heavily on the choice of the potential. The QM/MM description is more consistent: the vibrational free energy contribution is systematically unfavorable for the reaction ($\Delta G_{vib} > 0$), with the internal energy difference being negative and the entropic part positive.

As mentioned in the previous subsection, the effect of imposing the Eckart constraints on the absolute free energy values is a good measure for the accuracy of the Hessian. The differences between the values G_{vib} and G_{vib}^{Eck} are taken up in Table S1 of the Supporting Information and the average deviations in Table 8. It is found that QM Hessians are very sensitive to the Eckart constraints with an average shift in G_{vib} of 0.18 kcal/mol, while the sensitivity of QM/MM Hessians is noticeably better with an average shift of 0.02 kcal/mol. Note that these average deviations are based on the rather large REA, PROD1, and PROD2 molecules, since those have floppy modes with low-lying frequencies, which are absent in the smaller species. The different effect on QM and QM/MM Hessians can be explained, on one hand, by the geometry convergence of a QM system being delicate—since it depends on QM gradients which are sensitive to numerical integration errors themselves as well. On the other hand, the analytical QM second derivatives are particularly sensitive to the numerical integration accuracy of two electron integrals and convergence criteria of iterative loops (e.g., SCF loop, CPSCF loop). With 50 or more QM atoms in the QM description and only 11 or less atoms in the QM region of the QM/MM description, the QM/MM calculations are thus more accurate. These average deviations should be considered as errors inherent to the calculated data, originating from the present computational settings, just like experimental data having a limited accuracy imposed by the experimental setup.

The effect of the Eckart conditions on the reaction free energy can be seen in Tables 4−7 by comparing ΔG_{vib} with ΔG_{vib}^{Eck}. For QM Hessians, the deviation between FHVA and FHVA^{Eck} is on average about 0.11 kcal/mol and, for QM/MM Hessians, 0.02 kcal/mol. Not surprisingly, a deviation between FHVA and FHVA^{Eck} brings along a comparable deviation between MBH and MBH^{Eck}. The rather small deviations illustrate that the errors on the absolute vibrational free energies occasianally cancel out when taking the difference, but this behavior is not guaranteed. The expected accuracy of the free energy difference therefore must be on the same order as the accuracy of the absolute free energy values themselves, which is indeed the case.

Next, the influence of the use of MBH frequencies on the free energy difference is discussed, by comparing ΔG_{vib} [FHVA] with ΔG_{vib} [MBH]. The MBH approximation produces an average error of 0.09 and 0.04 kcal/mol when derived from QM and QM/MM Hessians, respectively, where the average is taken over all levels of theory and all block choices. The *absolute* free energies are drastically reduced by the MBH approach by over 145 kcal/mol when three blocks are used (data not shown); however, this significant shift is consistent between reactants and products such that it mostly cancels out when considering the free energy *differences* in eqs 32 and 33. This means that MBH alters the FHVA results relatively little. A comparison of ΔG_{vib}^{Eck} [FHVA] and ΔG_{vib}^{Eck} [MBH] shows similar errors of MBH^{Eck} with respect to FHVA^{Eck}. This significant cancellation of errors is in agreement with an earlier study on the reproduction of reaction rate constants with MBH by Ghysels et al.,[74] where errors canceled out in the difference $G(ts) - G(rea)$ between the transition state and the reactants. In the present study, the cancellation implies that the internal motions of the proposed blocks Ph, iP, and py are not crucial for the reactive behavior of the chemically active boron center, as expected. Indeed, the use of MBH has as much effect as have the Eckart constraints, which is a measure of the best accuracy that can be obtained with the given data (i.e., the Hessians).

When applied to transition state geometries, our approach can be used to estimate tunneling corrections, kinetic isotope effects, and local free energy estimates in the harmonic limit, without the cost of explicit conformational sampling. Indeed, for the estimates to be meaningful, only the frequency and the character of the lowest modes need be accurate. For instance, the kinetic isotope effect is closely related to the ratio of partition functions, which are governed by the low frequency modes. The studies in refs 25, 27, 37, 71, 74, and 75 have shown that MBH errors in partition functions, in free energy differences, and in reaction rates indeed cancel out when comparing two conformations if the same block choice is applied. Also, the application in the present paper supports the idea of the cancellation of errors. This makes MBH a promising method when it is applied on transition states, in particular for examining enzymatic reactions employing the QM/MM description.

The direct calculation of the QM/MM mobile block Hessian (case C, bottom part of tables) should yield the exact same frequencies as those based on the projection of the QM/MM full Hessian (case B, middle part of tables). The difference of 0.03 kcal/mol on average is indeed minimal and should be explained by numerical inaccuracies such as the finite convergence criteria in the CPSCF routine in Q-Chem and the numerical integration. In conclusion, the

new MBH implementation in the parallel Q-Chem/CHARMM interface is capable of reproducing the reference (FHVA) ΔG with satisfying accuracy.

Moreover, close inspection of the MBH values corresponding to one, two, or three blocks confirms the product rule as established in ref 74. In this paper, the correction in reaction rate due to the introduction of multiple blocks was found to be the product of the corrections due to the presence of each block individually. Similarly, the small deviations in free energy $\Delta\Delta G$ due to the introduction of single blocks approximately add up to the deviation in free energy due to multiple blocks, e.g., $\Delta\Delta G(\mathrm{MBH_{Ph}}) + \Delta\Delta G(\mathrm{MBH_{py}}) + \Delta\Delta G(\mathrm{MBH_{iP}}) \approx \Delta\Delta G(\mathrm{MBH_{Ph,\,py,\,iP}})$.

The variations in ΔG_{vib} induced by the choice of NMA model, more specifically the MBH, are negated by different errors. Indeed, the choice of a QM versus a QM/MM treatment shifts the FHVA values by 0.62 kcal/mol on average. The choice of the functional, on the other hand, is responsible for an average shift of 0.53 and 0.20 kcal/mol in the QM and QM/MM cases, respectively. These values lie higher than the typical errors encountered by MBH (0.10 for QM, 0.04 for QM/MM) such that MBH can be considered a minor source of deviations of the vibrational free energy difference, on the same order as the inherent accuracy of the data (0.10 for QM, 0.03 for QM/MM). Table 8 displays the hierarchy of the errors. The introduction of mobile blocks in the vibrational analysis has a significantly lower effect on ΔG compared to other computation parameters such as the functional and the QM versus QM/MM treatment. Therefore, the MBH approach not only is favorable because of the reduction in computational cost and memory requirements but is also capable of reproducing the full Hessian results with satisfactory to excellent accuracy.

V. Conclusion

The computation of vibrational frequencies from the analytical second derivatives matrix is a bottleneck in the hybrid QM/MM description due to the long-range Coulomb interactions when the electrostatic embedding scheme is employed. This is even the case for a small number of QM atoms, since additional CPSCF equations need to be solved for each MM atom displacement. Instead of using a cutoff technique that neglects interactions beyond a certain cutoff distance r_c, we introduce mobile blocks in the MM region. These blocks can translate/rotate as a whole, but their internal degrees of freedom are frozen in the vibrational analysis. Consequently, fewer CPSCF equations need to be solved, leading to a reduction in both computation time and memory requirements. In the chorismate mutase example, MBH decreases the CPU time to 54% of the full Hessian calculation, and the memory is reduced by a factor of roughly 5, when using mobile blocks of 15 atoms each. The computational profit further increases with increasing block size.

In this paper, the MBH formalism is established in view of the QM/MM interface between Q-Chem and CHARMM. A parallel version of MBH in the Q-Chem/CHARMM interface is now implemented in the latest version. Moreover, special attention is paid to the treatment of link atoms. The presence of link atoms creates artificial degrees of freedom which should be projected out of the Hessian in accordance with the definition of the total energy and the constraints imposed during the geometry optimization. Our suggestion is to impose the link atom to be located at a fixed scaled distance collinearly with the QM host atom and MM host atom. Formulas for the corresponding projection have been developed, and this projection is now available in the Q-Chem/CHARMM interface.

As an illustrative example, the vibrational free energy of bortezomib and the products after oxidative deboronation with the reagents H_2O_2 and methanol have been studied extensively, with four different levels of theory and a series of MBH block choices. Our results for this particular test system show an inherent error of 0.10 kcal/mol for the QM and 0.03 kcal/mol for the QM/MM vibrational free energy differences, which is quantified by imposing Eckart constraints. The introduction of mobile blocks introduces an error at a similar order of magnitude: 0.10 kcal/mol for QM and 0.04 kcal/mol for QM/MM vibrational free energy differences. Therefore, the considered block choices are reasonable approximations, especially given that much larger deviations are caused by the choice of functional (0.53 to 0.20 kcal/mol) or by the QM versus QM/MM description (0.62 kcal/mol). MBH is thus not only a computationally attractive method but also an adequate approximate approach for the calculation of thermodynamic quantities such as vibrational free energy differences.

Acknowledgment. This work is supported by the Fund for Scientific Research—Flanders (FWO), the Research Board of Ghent University (BOF), and BELSPO in the frame of IAP 6/27. This work is also supported by the Intramural Research Program of the National Heart, Lung and Blood Institute, National Institutes of Health (NIH). Funding was also received from the European Research Council under FP7 with ERC grant agreement number 240483. H.L.W. would like to acknowledge NIH (1K22HL088341-01A1) and the University of South Florida (start-up) for funding. Y.S. and J.K. would like to thank NIH for a Small Business Innovative Research grant (GM073408). Computational resources and services used in this work were provided by the Lobos cluster of the National Institutes of Health.

Supporting Information Available: The effect of imposing the Eckart constraints is a good measure for the accuracy of the Hessian; Table S1 contains the deviation of the vibrational free energy $\delta G_{\mathrm{vib}} = G_{\mathrm{vib}} - G_{\mathrm{vib}}^{\mathrm{Eck}}$ for the molecules REA, PROD1, and PROD2 calculated with the B3LYP, PBE, B3LYP-D, and PBE-D functionals and 6-311++G(d,p) basis set. It shows that QM Hessians are very sensitive to the Eckart constraints, while the sensitivity of the QM/MM Hessians is noticeably better.

This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Wilson, E. B.; Cross, P. C.; Decius, J. C. *Molecular Vibrations*; Dover Publications: New York, 1980.

(2) Fessenden, R. J.; Fessenden, J. S. *Organic chemistry*, 4th ed.; Brooks/Cole Publishing Company: Belmont, CA, 1990; pp 323−339.

(3) Cui, Q.; Bahar, I. *Normal Mode Analysis: Theory and applications to biological and chemical systems*; Chapman & Hall/CRC, Taylor & Francis Group: Boca Raton, FL, 2006; Mathematical and Computational Biology Series.

(4) Pulay, P. Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules. I. Theory. *Mol. Phys.* **1969**, *17*, 197.

(5) Pople, J. A.; Krisnhan, R.; Schlegel, H. B.; Binkley, J. S. Derivative studies in Hartree-Fock and Moller-Plesset theories. *Int. J. Quant. Chem.* **1979**, *Symp. 13*, 225–41.

(6) Saxe, P.; Yamaguchi, Y.; Schaefer, H. G. Analytic 2nd derivatives in restricted Hartree-fock theory - a method for hig-spin open-shell molecular wave-functions. *J. Chem. Phys.* **1982**, *77*, 5647–5954.

(7) Osamura, Y.; Yamaguchi, Y.; Saxe, P.; Vincent, M. A.; Gaw, J. F.; Schaefer, H. F. Unified theoretical treatment of analytic first and second energy derivatives in open-shell Hartree-Fock theory. *J. Chem. Phys.* **1982**, *72*, 131–139.

(8) Osamura, Y.; Yamaguchi, Y.; Saxe, P.; Fox, D. J.; Vincent, M. A.; Schaefer, H. F. Analytic 2nd derivative techniques for self-consistent-field wave-functions - a new approach to the solution of the coupled perturbed Hartree-Fock equations. *J. Mol. Struct.* **1983**, *103*, 183–196.

(9) Yamaguchi, Y.; Frisch, M. J.; Gaw, J.; Schaefer, H. F.; Binkley, J. S. Analytic evaluation and basis set dependence of intensities of infrared spectra. *J. Chem. Phys.* **1986**, *84*, 2262.

(10) Frisch, M. J.; Yamaguchi, Y.; Gaw, J.; Schaefer, H. F.; Binkley, J. S. Analytic Raman intensities from molecular electronic wave-functions. *J. Chem. Phys.* **1986**, *84*, 531.

(11) Frisch, M.; Head-Gordon, M.; Pople, J. Direct analytic SCF 2nd derivatives and electric-field properties. *Chem. Phys.* **1990**, *141*, 189–196.

(12) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions - Dielectric, Electrostatic and Steric Stabilization of Carbonium-Ion in Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.

(13) Singh, U. C.; Kollman, P. A. A Combined Abinitio Quantum-Mechanical and Molecular Me-chanical Method for Carrying out Simulations on Complex Molecular-Systems - Applications to the Ch3cl + Cl- Exchange-Reaction and Gas-Phase Protonation of Polyethers. *J. Comput. Chem.* **1986**, *7*, 718–730.

(14) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum-Mechanical and Molecular Mechanical Potential for Molecular-Dynamics Simulations. *J. Comput. Chem.* **1990**, *11*, 700–733.

(15) Lin, H.; Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(16) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. ONIOM: A multilayered integrated MO+MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)(3))(2)+H-2 oxidative addition. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(17) Dapprich, S.; Komaromi, I.; Byun, K.; Morokuma, K.; Frisch, M. J. A new ONIOM imple-mentation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *THEOCHEM* **1999**, *461*, 1–21.

(18) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.

(19) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(20) Woodcock, H. L.; Hodoscek, M.; Gilbert, A. T. B.; Gill, P. M. W.; Schaefer, H. F.; Brooks, B. R. Interfacing Q-Chem and CHARMM to perform QM/MM reaction path calculations. *J. Comput. Chem.* **2007**, *28*, 1485–1502.

(21) Gao, J.; Truhlar, D. G. Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

(22) Senn, H. M.; Thiel, W. QM/MM methods for biological systems. In *Atomistic Approaches in Modern Biology: from Quantum Chemistry to Molecular Simulations*; Springer-Verlag: Berlin, 2007; Vol. 268, pp 173−290.

(23) Vreven, T.; Byun, K. S.; Komromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2006**, *2*, 815.

(24) Cui, Q.; Karplus, M. Molecular properties from combined QM/MM methods. I. Analytical second derivative and vibrational calculations. *J. Chem. Phys.* **2000**, *112*, 1133–1149.

(25) Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Verstraelen, T.; Waroquier, M. Vibrational modes in partially optimized molecular systems. *J. Chem. Phys.* **2007**, *126*, 224102.

(26) Ghysels, A.; Van Neck, D.; Waroquier, M. Cartesian formulation of the Mobile Block Hessian approach to vibrational analysis in partially optimized systems. *J. Chem. Phys.* **2007**, *127*, 164108.

(27) Ghysels, A.; Van Speybroeck, V.; Pauwels, E.; Catak, S.; Brooks, B. R.; Van Neck, D.; Waroquier, M. Comparative study of various normal mode analysis techniques based on partial Hessians. *J. Comput. Chem.* **2010**, *31*, 994–1007.

(28) Currently described functionality is included in CHARMM version 36a3 and later and the current development version of Q-Chem (scheduled to be released as part of version 4.0).

(29) Adams, J.; Behnke, M.; Chen, S.; Cruickshank, A. A.; Dick, L. R.; Grenier, L.; Klunder, J. M.; Ma, Y. T.; Plamondon, L.; Stein, R. L. Potent and selective inhibitors of the proteasome: dipeptidyl boronic acids. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 333–338.

(30) Adams, J.; Kaufmann, M. Development of the proteasome inhibitor velcade (bortezomib). *Cancer Invest.* **2004**, *22*, 304–311.

(31) Goldstone, J.; Salam, A.; Weinberg, S. Broken symmetries. *Phys. Rev.* **1962**, *127*, 965–970.

(32) Angyan, J. G. Wigner's (2n+1) rule for nonlinear Schrodinger equations. *J. Math. Chem.* **2009**, *46*, 1–14.

(33) Head-Gordon, M.; Pople, J. A. Optimization of wave function and geometry in the finite basis hartree-fock method. *J. Phys. Chem.* **1988**, *92*, 3063–3069.

(34) Ochsenfeld, C.; Head-Gordon, M. A reformulation of the coupled perturbed self-consistent field equations entirely within a local atomic orbital density matrix-based scheme. *Chem. Phys. Lett.* **1997**, *270*, 399–405.

(35) Woodcock, H. L.; Zheng, W. J.; Ghysels, A.; Shao, Y. H.; Kong, J.; Brooks, B. R. Vibrational subsystem analysis: A method for probing free energies and correlations in the harmonic limit. *J. Chem. Phys.* **2008**, *129*, 214109.

(36) Liang, W.; Zhao, Y.; Head-Gordon, M. An efficient approach for self-consistent-field energy and energy second derivatives in the atomic-orbital basis. *J. Chem. Phys.* **2005**, *123*, 194106.

(37) Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Brooks, B. R.; Waroquier, M. Normal modes for large molecules with arbitrary link constraints in the Mobile Block Hessian approach. *J. Chem. Phys.* **2009**, *130*, 084107.

(38) Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodoscek, M.; Brooks, B. R. Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method. *J. Chem. Phys.* **2002**, *117*, 10534–10547.

(39) Li, H.; Jensen, J. H. Partial Hessian vibrational analysis: the localization of the molecular vibrational energy and entropy. *Theor. Chem. Acc.* **2002**, *107*, 211–219.

(40) Jin, S. Q.; Head, J. D. Theoretical Investigation of Molecular Water-Adsorption on the Al(111) Surface. *Surf. Sci.* **1994**, *318*, 204–216.

(41) Calvin, M. D.; Head, J. D.; Jin, S. Q. Theoretically modelling the water bilayer on the Al(111) surface using cluster calculations. *Surf. Sci.* **1996**, *345*, 161–172.

(42) Head, J. D. Computation of vibrational frequencies for adsorbates on surfaces. *Int. J. Quantum Chem.* **1997**, *65*, 827–838.

(43) Head, J. D.; Shi, Y. Characterization of Fermi resonances in adsorbate vibrational spectra using cluster calculations: Methoxy adsorption on Al(111) and Cu(111). *Int. J. Quantum Chem.* **1999**, *75*, 815–820.

(44) Head, J. D. A vibrational analysis with Fermi resonances for methoxy adsorption on Cu(111) using ab initio cluster calculations. *Int. J. Quantum Chem.* **2000**, *77*, 350–357.

(45) Besley, N. A.; Metcalf, K. A. Computation of the amide I band of polypeptides and proteins using a partial Hessian approach. *J. Chem. Phys.* **2007**, *126*, 035101.

(46) Derat, E.; Bouquanta, J.; Humbel, S. On the link atom distance in the ONIOM scheme. An harmonic approximation analysis. *THEOCHEM* **2003**, *632*, 61–69.

(47) MacKerel, A., Jr.; Brooks, C., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*; Schleyer, v. R. et al., Eds.; John Wiley & Sons: Chichester, U. K., 1998; Vol. 1, pp 271−277.

(48) Bakowies, D.; Thiel, W. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

(49) Swart, M. AddRemove: A new link model for use in QM/MM studies. *Int. J. Quantum Chem.* **2003**, *91*, 177–183.

(50) CP2K Developers Home Page. http://cp2k.berlios.de (accessed October 15, 2010).

(51) Korambath, P. P.; Kong, J.; Furlani, T. R.; Head-Gordon, M. Parallelization of analytical Hartree-Fock and density functional theory Hessian calculations. Part I: parallelization of coupled-perturbed Hartree-Fock equations. *Mol. Phys.* **2002**, *100*, 1755–1761.

(52) Woodcock, H. L.; Hodoscek, M.; Sherwood, P.; Lee, Y. S.; Schaefer, H. F.; Brooks, B. R. Exploring the quantum mechanical/molecular mechanical replica path method: a pathway optimization of the chorismate to prephenate Claisen rearrangement catalyzed by chorismate mutase. *Theor. Chem. Acc.* **2003**, *109*, 140–148.

(53) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R., Jr.; Evanseck, J. D.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics Studies of proteins. *J. Phys. Chem.* **1998**, *102*, 3586–3616.

(54) MacKerell, A. D., Jr; Feig, M.; Brooks, C. L., III. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(55) Fahy, B. N.; Schlieman, M. G.; Virudachalam, S.; Bold, R. J. Schedule-dependent molecular effects of the proteasome inhibitor bortezomib and gemcitabine in pancreatic cancer. *J. Surg. Res.* **2003**, *113*, 88–95.

(56) Nawrocki, S. T.; Carew, J. S.; Pino, M. S.; Highshaw, R. A.; Andtbacka, R. H. I.; Dun-ner, K.; Pal, A.; Bornmann, W. G.; Chiao, P. J.; Huang, P.; Xiong, H.; Abbruzzese, J. L.; McConkey, D. J. Aggresome disruption: a novel strategy to enhance bortezomib-induced apoptosis in pancreatic cancer cells. *Cancer Res.* **2006**, *66*, 3773–3781.

(57) Nawrocki, S. T.; Carew, J. S.; Pino, M. S.; Highshaw, R. A.; Dunner, K.; Huang, P.; Abbruzzese, J. L.; McConkey, D. J. Bortezomib borate anion in a hydrogen-bonded host lattice sensitizes pancreatic cancer cells to endoplasmic reticulum stress-mediated apoptosis. *Cancer Res.* **2005**, *65*, 11658–11666.

(58) McCormack, T.; Baumeister, W.; Grenier, L.; Moomaw, C.; Plamondon, L.; Pramanik, B.; Slaughter, C.; Soucy, F.; Stein, R. L.; Zuhl, G.; Dick, L. R. Active site-directed inhibitors of phodococcus 20 S proteasome: kinetics and mechanism. *J. Biol. Chem.* **1997**, *272*, 26103–26109.

(59) Pekol, T.; Daniels, J. S.; Labutti, J.; Parsons, I.; Nix, D.; Baronas, E.; Hsieh, F.; Gan, L.-S.; Miwa, G. Human metabolism of the proteasome inhibitor bortezomib: identication of circulating metabolites. *Drug Metab. Dispos.* **2005**, *33*, 771–777.

(60) Labutti, J.; Pearsons, I.; Huang, R.; Miwa, G.; Gan, L.-S.; Daniels, J. S. Oxidative deboronation of the peptide boronic acid proteosome inhibitor bortezomib: contributions from reactive oxygen species in this novel cytochrome P450 reaction. *Chem. Res. Toxicol.* **2006**, *19*, 539–546.

(61) Larkin, J. D.; Markham, G. D.; Milkevitch, M.; Brooks, B. R.; Bock, C. W. Computational Investigation of the Oxidative Deboronation of Boroglycine, H_2N-CH_2-$B(OH)_2$, Using H_2O and H_2O_2. *J. Phys. Chem. A* **2009**, *113*, 11028–11034.

(62) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic be-havior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(63) Lee, C. T.; Yang, W. T.; 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.

(64) Ernzerhof, M.; Perdew, J. P.; Burke, K. Coupling-constant dependence of atomization energies. *Int. J. Quantum Chem.* **1997**, *64*, 285.

(65) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew-Burke-Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

(66) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable pa-rameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–69.

(67) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dis-persion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(68) Feyereisen, M.; Fitzegerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(69) Woon, D. E.; Dunning, T. H. J. Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **1994**, *100*, 2975–2988.

(70) Vangunsteren, W. F.; Berendsen, H. J. C. Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol. Phys.* **1977**, *34*, 1311–1327.

(71) Ghysels, A.; Verstraelen, T.; Hemelsoet, K.; Van Speybroeck, V.; Waroquier, M. TAMkin: a versatile package for vibrational analysis and chemical kinetics. *J. Chem. Inf. Model.* **2010**, 1736–1750.

(72) Grochowski, P. Rotational symmetry of the molecular potential energy in the Cartesian coordinates. *Theor. Chem. Acc.* **2008**, *121*, 257–266.

(73) Brandhorst, K.; Grunenberg, J. Efficient computation of compliance matrices in redundant internal coordinates from Cartesian Hessians for nonstationary points. *J. Chem. Phys.* **2010**, *132*, 184101.

(74) Ghysels, A.; Van Speybroeck, V.; Verstraelen, T.; Van Neck, D.; Waroquier, M. Calculating reaction rates with partial Hessians: Validation of the mobile block Hessian approach. *J. Chem. Theory Comput.* **2008**, *4*, 614–625.

(75) Ghysels, A.; Van Speybroeck, V.; Pauwels, E.; Van Neck, D.; Brooks, B. R.; Waroquier, M. Mobile Block Hessian approach with adjoined blocks: an efficient approach for the calculation of frequencies in macromolecules. *J. Chem. Theory Comput.* **2009**, *5*, 12031215.

(76) Mc Quarrie, D. A.; Simon, J. D. *Physical Chemistry - a molecular approach*; University Science Books: Sausalito, CA, 1997; pp 1075−1079.

(77) Brooks, B. R.; Janezic, D.; Karplus, M. Harmonic-Analysis of Large Systems. 1. Methodology. *J. Comput. Chem.* **1995**, *16*, 1522–1542.

(78) Janezic, D.; Brooks, B. R. Harmonic-Analysis of Large Systems. 2. Comparison of Different Protein Models. *J. Comput. Chem.* **1995**, *16*, 1543–1553.

(79) Janezic, D.; Venable, R. M.; Brooks, B. R. Harmonic-Analysis of Large Systems. 3. Comparison with Molecular-Dynamics. *J. Comput. Chem.* **1995**, *16*, 1554–1566.

CT100473F

JCTC Journal of Chemical Theory and Computation

Configurational Entropy Reallocation and Complex Loop Dynamics of the Mosquito-Stage Pvs25 Protein Complexed with the Fab Fragment of the Malaria Transmission Blocking Antibody 2A8

Athanassios Stavrakoudis*,† and Ioannis G. Tsoulos‡

Department of Economics, University of Ioannina, Ioannina, Greece, and Department of Communications, Informatics & Management, Technical Educational Institute of Epirus, Arta, Greece

Received September 22, 2010

Abstract: Pvs25 is a protein of unique 3D structure, and it is characterized by the presence of repeated EGF-like domains and 11 disulfide bonds. It is a very important candidate for the transmission-blocking malaria vaccine, as it plays an important role in mosquito infection by Plasmodium parasites. Recently, the X-ray structure of the protein complexed with the transmission blocking antibody 2A8 has been reported. In this study, we report the loop reorganization of the Pvs25 protein based on configurational entropy calculations and dihedral principal component analysis as revealed from the protein complex and free molecular dynamics simulations. While the total entropy of the protein was estimated to be almost the same in the free and complex trajectories, the partition of the entropy contribution in the loop fragments of the protein revealed interesting entropy reallocation after the 2A8 antibody binding. Interestingly, the 51−71 protein loop experienced a significant reduction in its configurational entropy, while other parts of the protein did not show any difference in it, or even showed an entropy increase. This trend in entropy redistribution was found to be in direct relationship with specific interactions with the antibody's binding site. Results from root-mean-square fluctuations/deviations and dihedral angle principal component analysis further support this finding.

1. Introduction

Pvs25 is a protein derived from the malaria parasite Plasmodium,[1,2] a worldwide spread parasite.[3,4] The protein plays an essential role in infecting mosquitoes and thus transmitting malaria.[5] It is an important target in the development of a vaccine.[6,7] The 3D structure of the Pvs25 protein has been solved[8] by X-ray and is characterized by the presence of 11 disulfide bonds (Figure 1). This high number of disulfide bonds makes Pvs25 a protein of unique structure. For example, a recent review about the classification of disulfide bonds in proteins analyzed the existence of up to 10 disulfide bonds in proteins.[9] Given the high degree of interest in blocking the malaria transmission by mosquitoes, it is interesting to explore the dynamics of protein/antibody binding.

Computer simulation of molecular dynamics is a well established method for studying several aspects of biomolecular structure and function.[10–12] Moreover, biomolecular modeling can complement experimental studies,[13] and recent studies have been used in order to elucidate the dynamics of protein folding,[14] to explore the immunogenicity of peptide−vaccine candidates,[15] to facilitate vaccine design[16] to help in rational drug design,[17] to account for the peptide's flexibility in immunological complexes,[18] or to dock molecules into binding sites of proteins.[19] It has also been argued

* To whom correspondence should be addressed. E-mail: astavrak@cc.uoi.gr.

† University of Ioannina.

‡ Technical Educational Institute of Epirus.

10.1021/ct100543c
Published on Web 12/29/2010

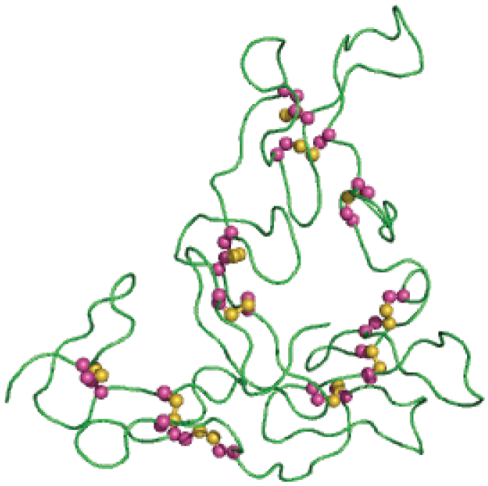


Figure 1. Ribbon representation of the X-ray structure of the Pvs25 protein. Disulfide bonds are highlighted with balls and sticks. C^{α} and C^{β} atoms of Cys residues are colored in magenta, while S^{γ} atoms are colored in yellow.

that molecular dynamics simulations can be used as an aiding tool in experimental studies.[13]

Biomolecular plasticity affects in an essential way many biological functions.[20,21] The advancements in computer power and statistical mechanics methods have contributed a lot to targeting biomolecular flexibility.[22] In this study, computer simulation molecular dynamics have been performed on the Pvs25 protein, and its complex with the Fab fragment of the 2A8 Fab antibody, in order to get insight into the binding mode. We have found that not all 11 disulfide loops of the Pvs25 protein behave in the same way. Differences in loop dynamics are directly related with protein/antibody contacts. The dynamics of the fourth loop, formed by Cys residues 51 and 71, are most profoundly affected. This makes the 51−71 region a candidate for engineered protein/antibody interactions in targeting the blocking of malaria transmission by mosquitoes.

2. Methods

2.1. Setup of the System and MD Simulations. Initial coordinates of the Pvs25 and the Fab fragment of malaria transmission blocking antibody 2A8 were downloaded from the Protein Data Bank,[23] PDB code: 1z3g.[8] The protonation status of histidine side chains was estimated with the REDUCE program.[24] Topology and force field parameters for all atoms were assigned from the CHARMM22-CMAP parameter set.[25,26] It has been found that the addition of cross terms with the CMAP potential improves system parametrization and helps to avoid undesired helical transitions.[27,28] Hydrogen atoms were added with the VMD program[29] and its autopsf utility. The antibody/protein complex was centered in a rectangular box with dimensions 141.72 × 87.89 × 109.70 $Å^3$. The box was filled with 24 429 TIP3P water molecules and neutralized with the addition of 40 Na^+ and 33 Cl^- ions to approximate a physiological ionic concentration of 0.1 mM. The total number of atoms of the whole system was 126 818. Nonboned van der Waals interactions were gradually turned off at a distance between 12 and 14 Å. Long-range electrostatics were calculated with the PME method.[30] Nonbonded forces and PME electrostatics were computed every second step. The pair list was updated every 10 steps. Bonds to hydrogen atoms were constrained with the SHAKE method,[31] allowing a 2 fs time step for integration. The system was initially subjected to energy minimization with 5000 steps. The temperature of the system was then gradually increased to 310 K, with Langevin dynamics using the NVT ensemble, during a period of 3000 steps, by stepwise reassignment of velocities every 500 steps. The simulation was continued at 310 K for 100 000 steps (200 ps). During the minimization and equilibration phases, protein backbone atoms (N, C^{α}, C′, O) were restrained to their initial positions with a force constant of 50 kcal mol^{-1} $Å^{-2}$. The system was equilibrated for another 200 ps with the force constant reduced to 50 kcal mol^{-1} $Å^{-2}$. Finally, 400 ps of NVT simulation at 310 K were performed with total elimination of the positional restraints. The simulation was passed to the productive phase, by applying constant pressure with the Langevin piston method.[32] Pressure was maintained at 1 atm and a temperature of 310 K. The results are based on a period of 20 ns of this isothermal−isobaric (NPT) run. Snapshots were saved to disk at 1 ps intervals for structural analysis. Results from this trajectory are denoted as the complex trajectory for the rest of this article.

An identical protocol was followed for the antibody-free protein (PDB code 1z27) to obtain the free trajectory of the protein.

Trajectory analysis was performed with Eucb[33] and Carma[34] software packages. Appropriate corrections have taken into account dealing with circular data statistics.[35] Hydrogen bonds were estimated with a geometrical criterion as described elsewhere.[36] Structural figures were prepared with PyMOL (www.pymol.prg).

2.2. Dihedral Angle Principal Component Analysis. Principal component analysis (PCA) is a standard method for analyzing MD trajectories, where the reduction of the dimensionality of a high-dimensional data set is desired.[37] The dihedral based PCA (dPCA) has been applied[38−40] to explore the energy landscape of a biomolecule. Calculations of dPCA have been performed with Carma.[34]

2.3. Entropy Calculations. Molecular dynamics simulations offer various methods to estimate the absolute or relative entropy.[41,42] Schliter's formulation[43] was used for the estimation of the configurational entropy:

$$S_{\text{true}} = \frac{1}{2}k_{\text{B}} \ln \det\left[1 + \frac{k_{\text{B}}Te^2}{\hbar^2}\mathbf{M}\sigma\right] \qquad (1)$$

where S is an upper estimation of the true entropy (S_{true}), k_{B} is Boltzmann's constant, T is the absolute temperature (in which the system was simulated), e is the Euler number, $\hbar$ is the Plank constant divided by 2π, $\mathbf{M}$ is the mass matrix that holds on the diagonal the masses belonging to the atomic Cartesian degrees of freedom, and σ is the covariance matrix of atom positional fluctuations:

$$\sigma_{ij} = \langle x_i - \langle x \rangle \rangle \langle y_i - \langle y \rangle \rangle \qquad (2)$$

Entropy calculations were performed with the backbone atoms (N, C^{α}, C′) of the peptide from the bound and free

trajectories respectively, at 0.1 ns intervals (100 frames). Two separate trajectories (for example, free and complex trajectories of a peptide) can be combined. Thus one trajectory can be appended at the end of the other trajectory, and the plot of configurational entropy S against time can be used as an assessment of the overlap between configurational spaces sampled in the two simulations.[44] Such trajectories have been derived for the backbone (bb) atoms (N, C^{α}, C′) of the peptide from the last 10 ns of the free (f) and bound (b) trajectories. Both appending sequences were applied, resulting in $S_{bb}^{trA+trB}$ and $S_{bb}^{trB+trA}$ calculations, where the trB trajectory was appended to the trA one (trA+trB) or the trA trajectory was appended to the trB one (trB+trA). Plotting the calculated values of S from both the combined trajectories over time demonstrates the relative size and overlap of sampled trajectories. Plotting S over time after the combination of two trajectories results in three cases,[44] briefly described as follows:

1 S increases after appending one trajectory to the other, with a jump observed at this point. Thus, the two trajectories do not overlap, or there is only a small overlap between them.
2 S evolves smoothly after the appending of the trajectories, without an observable perturbation of the line of S over time; thus, the two trajectories show significant overlap.
3 The S curve increases during the time of the first trajectory but decreases a little after the appending of the second trajectory; thus, the second trajectory samples a smaller configurational space than the first one, which also contains the configurational space visited by the second one.

The calculation of entropy buildup curves has been performed with the Eucb program,[33] which utilizes a routine adapted from numerical recipes[45] for the computation of the eigenvalues. Moreover, trajectories for the backbone, heavy, or heavy side chain atoms have been extracted from the complex and free trajectories for each residue of the Pvs25 sequence, in order to obtain the entropy difference per residue.[46]

2.4. Buried Surface Area. Calculation of the buried surface area (BSA) was performed with the NACCESS program,[47] based on the formula

$$\mathrm{BSA} = S_p + S_a - S_c \quad (3)$$

Thus, the BSA is the difference of the surface accessible area of the complex (S_c) from the sum of the surface accessible areas of the protein (S_p) and antibody molecule (S_a). Calculations were performed for all frames of the complex trajectory in order to get and characterize the times series of BSA.

3. Results and Discussion

3.1. Backbone Dynamics of the Protein. The Pvs25 protein and its 2A8 antibody complex remained stable during both free and complex MD trajectories. Figure 2 shows the root-mean-square fluctuations (RMSF) of C^{α} atoms and deviations (RMSD) of backbone (N, C^{α}, C′) atoms of the protein and antibody's heavy (H) and light chains (L).

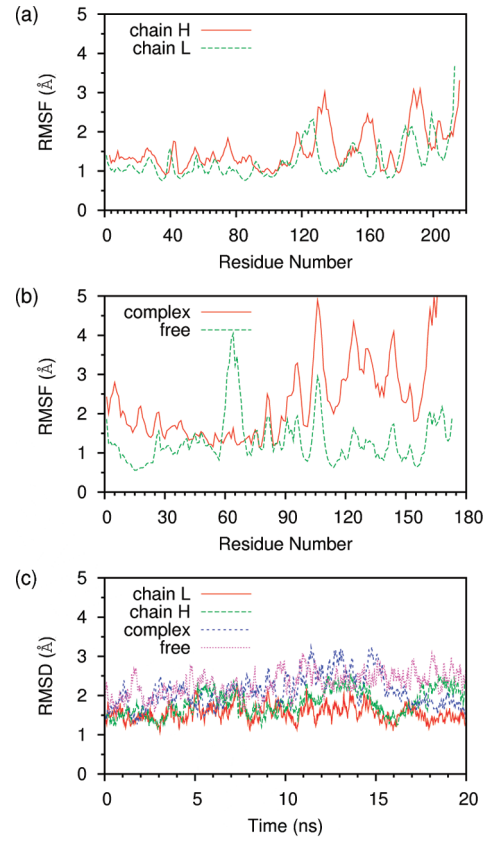


Figure 2. Root mean square fluctuation (RMSF) of Cα atoms and root-mean-square deviation (RMSD) time series of backbone atoms (N, C^{α}, C′) after fitting the corresponding atom positions from MD trajectories to the initial (X-ray) coordinates. (a) RMSF of C^{α} atoms of heavy (H) and light (L) chains of the antibody. (b) RMSF of C^{α} atoms of the protein in complex and free forms. (c) RMSD of backbone atoms of heavy (H) and light (L) chains of the antibody, and the backbone atoms of the protein in complex and free forms.

As revealed in part c of Figure 2, the RMSD of the backbone atoms of all protein chains fluctuated between 1 and 3 Å during the simulation time, without any significant breaks or jumps. Interestingly, the Pvs25 protein showed greater mobility in comparison to antibody's heavy or light chains. The mobility of the protein's backbone atoms was found to be approximately the same in free and complex trajectories, where the average values for the RMSD (with std. dev.) were 2.3 (0.3) Å and 2.1 (0.4) Å, respectively.

An interesting feature is revealed in part b of Figure 2, where the RMSF values of the protein's C^{α} atoms is displayed. The first 40 residues of the Pvs25 protein (N-terminal) showed more flexibility in the complex form. This picture was inversed for the next 40 residues, where comparable values were observed, or at residues around Glu_{59} (peak of the RMSF line of protein's free trajectory), a significant reduction of the protein's C^{α} atom's mobility was recorded. The C-terminal half of the protein showed surprisingly high mobility in the complex form. The trend of the RMSF values was very similar in the free and complex trajectories; however, the calculated values were higher in the complex than the free trajectory. In general, it is admitted that binding reduces a protein's mobility. However, in this

case, it is evident that this happened selectively in some part of the protein, while the remaining part increased its mobility.

The Pvs25 protein contains 22 Cys residues that are all paired in 11 disulfide bonds. Pvs25 is a unique protein in this respect, especially if its 177 residue length is taken into consideration. Its unique 3D structure is characterized by the presence of repeated EGF like domains.[8] In order to see how the mobility of these disulfide loops is affected by the 2A8 binding of the Pvs25 protein, we split the protein's sequence into 11 (overlapping) parts, according to disulfide bond formation, and we measured the RMSD of the backbone atoms of these loops. The results of this procedure are displayed in Figure 3. The first three loops, 8−22, 24−36, and 42−57, did not show any difference worth mentioning in the time evolution of the RMSD. The same conclusion can be drawn for the last seven loops, 73−84, 89−100, 94−113, 115−129, 137−148, 141−157, and 159−172. The RMSD values could be very small, as in loops 8−22 and 115−129, or bigger like in loops 94−113 or 159−172, but in all of these cases, the pattern of the RMSD time evolution was very similar. This fact indicates that the antibody binding of the protein greatly influenced the intrinsic mobility of the protein's loops.

The most notable difference between the free and complex trajectories concerns the 51−71 loop. It must be underlined also that this loop contains the Glu_{59} residue, which showed the highest RMSF value in the free trajectory (Figure 2). Thus, the 0.5 Å value that was approximately observed during the complex trajectory increased by approximately 1−1.2 Å in the free trajectory. The 51−71 disulfide loop directly contacts the antibody's heavy chain, and its reduced mobility after binding is somewhat expected. What is most notable here is that this is the only loop that is affected by the binding.

3.2. Dihedral Principal Component Analysis. In order to explore more thoroughly the backbone dynamics of the protein loops and the influence of the binding of the 2A8 antibody, we performed a principal component analysis of the complex and free trajectories, based on backbone dihedral angles. This technique has been routinely used during recent years in energy landscape studies of peptides and small proteins.[38,39,48] At this stage, as in the RMSD calculation of the disulfide loops, we examined the energy landscape of the 11 loops of the protein, as determined by the presence of 11 disulfide bonds. The results of these calculations are illustrated in Figure 4.

Loop 8−22 showed almost identical results before and after the binding. This is also implied from the RMSD analysis, previously analyzed (Figure 3).

Loop 24−36 showed an interesting feature because of the widening of distribution in the dihedral angle space. While in the free trajectory the protein's structures clustered mainly in very close conformations, in the complex structure, a more wide distribution can be observed. The inverse picture can be seen for the 42−57 loop, where we observed a reduction in conformational space sampling upon antibody binding of the protein.

Similarly to the situation in the 34−46 loop, a significant reduction in the sampled conformational space can be extracted for the 51−71 loop, from the corresponding parts of Figure 4. A second small conformational cluster is also seen in the complex trajectory. However, it seems that the antibody binding of the Pvs25 protein restricted the mobility on this fragment. The main big cluster in the complex trajectory has a center very close to those observed in the free trajectory, so it can be concluded that the 51−71 loop

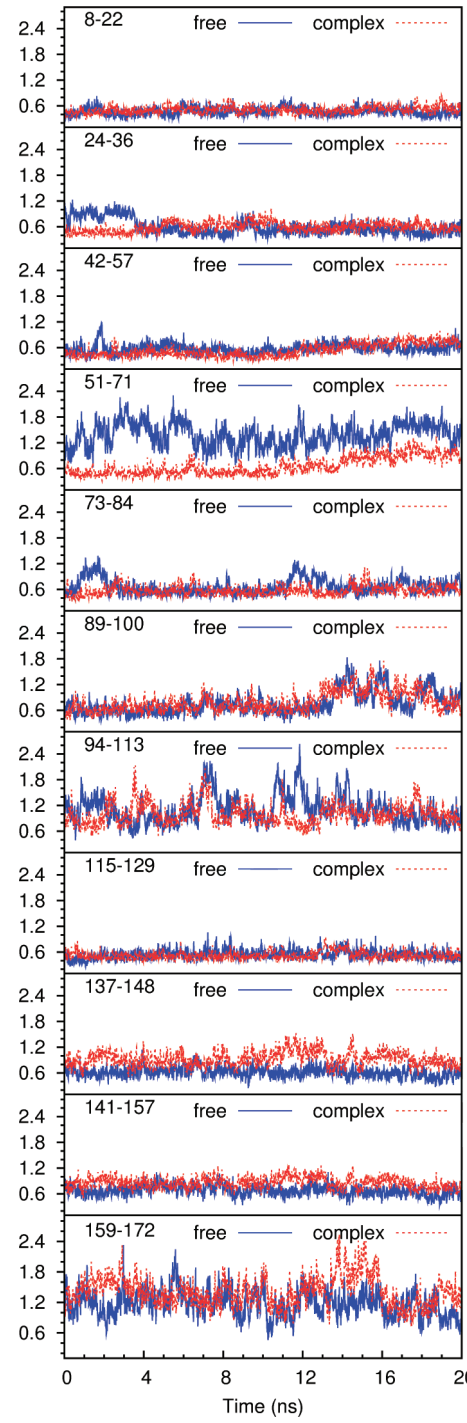


Figure 3. Time evolution of root-mean-square deviation (RMSD) of the backbone atoms of the 11 disulfide loops of the Pvs25 protein, in free and complex trajectories. RMSD is measured in Ångstroms. Numbers, e.g., 8−22, on the top left side of the plots indicate the residues that form the corresponding disulfide bond.

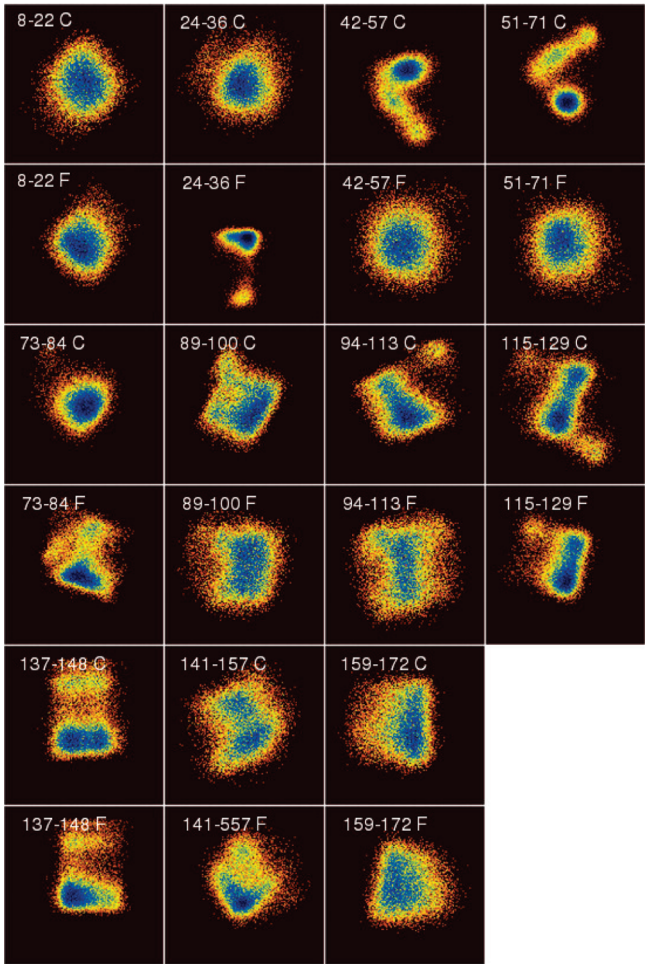


Figure 4. Dihedral principal component analysis of the 11 loops of the Pvs25 by disulfide bonds. Numbers in the plots indicate the cysteine residues that form the disulfide bonds. The letter "F" or "C" is used in order to discriminate the free or complex MD trajectory, respectively. All diagrams shown in this figure are pseudocolor representations of density functions corresponding to the projections of the fluctuations of the backbone dihedral angles (φ,Ψ) on the planes of the top two eigenvectors. The density function shown is $\triangle G = -k_B T \ln(p/p_{max})$ where k_B is the Boltzmann constant, T is the temperature in Kelvin, and p and p_{max} are probabilities obtained from the distribution of the principal components for each structure (frame) from the corresponding trajectory. The $\triangle G$ values obtained from this procedure are on an arbitrary scale in the sense that they depend on the binning procedure used for calculating the p and p_{max} values. For all diagrams of this figure, the raw data were binned on a square matrix of size $N/2$, where N is the number of frames of the corresponding trajectory.

sampled, in the complex trajectory, a part of the conformational space sampled in the free trajectory.

The rest of the loops showed remarkable similarity in dihedral principal component analysis results, before and after the antibody binding.

Quite encouraging for the MD analysis process, the results from the dPCA and loop RMSD fall in line and support the hypothesis that the flexibility in the protein's backbone, after the antibody binding, followed complex dynamics, without being uniformly distributed. Evidence is now accumulated that the flexibility of 51−71 loop, which dominates the protein/antibody interaction interface, was considerably reduced upon antibody binding. At the same time, other parts of the protein retained almost the same flexibility, or even showed an increase of their flexibility upon binding.

3.3. Configurational Entropy Analysis. The reduction of the configurational entropy of charged residues involved in protein/protein interaction interfaces is a well-known issue.[46]

We have extracted the heavy atoms of the protein from both the complex and free trajectories, and we have calculated the configurational entropies. The values we obtained were 36.31 and 36.21 kJ K^{-1} mol^{-1}, respectively. We also measured the configurational entropy of the backbone atoms of the protein, and we obtained values of 11.99 and 11.61 kJ K^{-1} mol^{-1} for the complex and free trajectories, respectively. The negligible difference in a protein's configurational entropy upon antibody complexation is quite interesting and deserves further investigation.

We split the protein sequence into 11 parts, as defined by the protein's disulfide bonds, and we extracted the coordinates of the backbone atoms from both trajectories. We appended the complex trajectory to the free trajectory (and *vice versa*) for all 11 loops, and we applied a positional least-squares fitting of the combined trajectory frames' coordinates to the first frame to remove any translational/rotational components in the configurational entropy calculations. We then calculated the configurational entropy of the combined trajectories at a 0.5 ns time interval. Then, we plotted the buildup entropy curves over time. The results of this procedure are illustrated in Figure 5. From a visual inspection of these plots, one can estimate if the conformational space sampled in the two trajectories overlapped or not, or if the two trajectories sampled different conformational spaces.

It can be easily extracted from Figure 5 that not all disulfide loops behave in the same way during the complex and free trajectories. For example, the configurational entropy of some loops located at the middle of the protein's sequence (42−57, 51−71, 73−84, or 89−100) experienced reduced values upon complexation, while some other loops located at the C-terminal part of the protein's sequence (137−148, 141−157, or 159−172) showed increased values upon complexation. Thus, it seems that despite the conservation of the protein's configurational entropy upon complexation, the protein did not remain static but redistributed its flexibility in order to adapt to conformational changes imposed by the antibody binding. This observation is also in line with the RMSF calculations analyzed previously (Figure 2) and dpCA calculations (Figure 9).

The first loop, 8−22, did not show any difference in the configurational entropy upon binding. The calculated ΔS was only −0.4 J K^{-1} mol^{-1}. This is in perfect agreement with the dPCA calculations of the 8−22 loop. The loops 24−36 and 42−57 showed a moderate decrease of −10.8 and −9.8 J K^{-1} $mol^{-1,}$ respectively, in the configurational entropy of the backbone atoms.

In line with the RMS (Figure 2) and dPCA (Figure 4) analysis, loop 51−71 experienced a great decrease in configurational entropy of −54.3 J K^{-1} mol^{-1}. Entropy

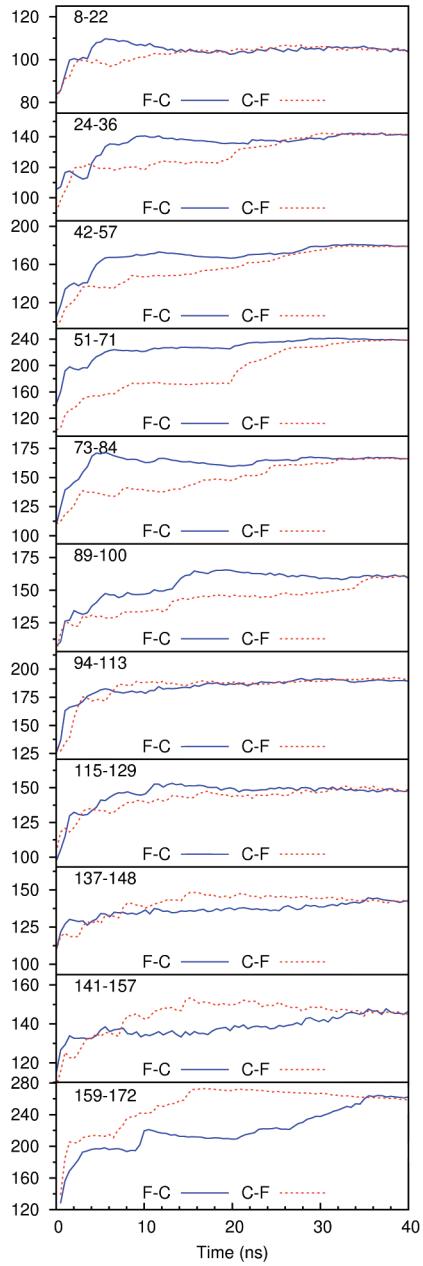


Figure 5. Configurational entropy of the 11 disulfide loops. Entropy units in the vertical axis of the plots are J K^{-1} mol^{-1}. Covariance matrices were generated after a least-squares fitting of atoms' positions of the protein to initial (X-ray) coordinates. The configurational entropy buildup curve was calculated every 500 frames (0.5 ns). The "F−C" notation indicates the appending of the complex to the free trajectory, while the "C−F" notation indicates the appending of the free to the complex trajectory.

calculations of the backbone atoms of this loop further confirmed the hypothesis that antibody binding of the Pvs25 protein greatly reduced the flexibility of this protein fragment.

Moderate entropy decreases have been observed in the 73−84 and 79−100 loops, with ΔS values of −13.1 and −19.9 J K^{-1} mol^{-1}, respectively. The values are comparable with those observed for loops 24−36 and 42−57, preceding loop 51−71. As in the first loop, 8−22, the ΔS value for the 94−113 loop was found to be quite close to 0, a value of 1.1 J K^{-1} mol^{-1}. If we consider loop 51−71 at site 0,

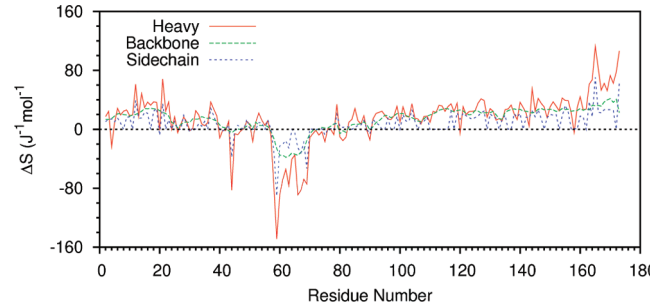


Figure 6. Differences in intraresidue configurational entropy between the complex and free trajectories measure the heavy, backbone, or side chain heavy atoms.

then loops at position ±1,2 showed a moderate decrease in ΔS and loops at position ±1,2 showed almost no difference in ΔS values upon complexation. This trend fits well with the hypothesis that the conformational flexibility restriction focused on the 51−71 loop and died out as we drew away. The trend of configurational entropy reduction fits also very well with the calculated RMSF values (Figure 2).

The remaining part of the protein showed mixed results about the ΔS. Most interesting is the increase of configurational entropies of the last two loops, 141−157 and 159−172, by 11.7 and 31.3 J K^{-1} mol^{-1}, respectively. The redistribution of configurational entropy[49,50] is an important feature of the protein/antibody binding studied here.

Additional insight into the role of the configurational entropy in protein/antibody binding can be provided by examination of the per residue contribution of the configurational entropy. Figure 6 shows the intraresidue contribution in ΔS of the backbone, heavy, and heavy side chain atoms. Similarly with other analyzed observations, residues of loop 51−71 contributed with highly negative values to ΔS. Interestingly, residues of the first two loops 8−22 and 24−36 showed mostly positive ΔS values. The same conclusion can be drawn for the C-terminal part of the peptide, in line with the backbone entropy per loop difference analyzed in the preceding paragraphs. The most negative peaks of the heavy atom line correspond to residues Lys_{44}, Glu_{59}, and Val_{66}. Unsurprisingly, these residues made significant contacts with the antibody, and the reduction of their flexibility upon binding is highly expected.

The negative valley of ΔS (Figure 6) for the backbone atoms corresponds to the 58−72 sequence of the protein, which consists mostly of the 51−71 loop. There is also a small region, 43−47, of marginally negative values in the backbone entropy per residue. These findings are in excellent agreement with other parts of the MD analysis presented in this work and further corroborate the hypothesis of entropy redistribution in the binding of the Pvs25 protein by the 2A8 antibody.

The difference in the protein's configurational entropy upon antibody binding has been visualized in Figure 7. From this representation, it can be seen that region 58−69, which constituted the main region of the binding site, showed the greatest entropy reduction upon binding (colored with blue). The C-terminal end of the protein, some 45−50 Å away from the binding site, showed a considerable increase (colored red) in configurational entropy upon binding. In general, the color

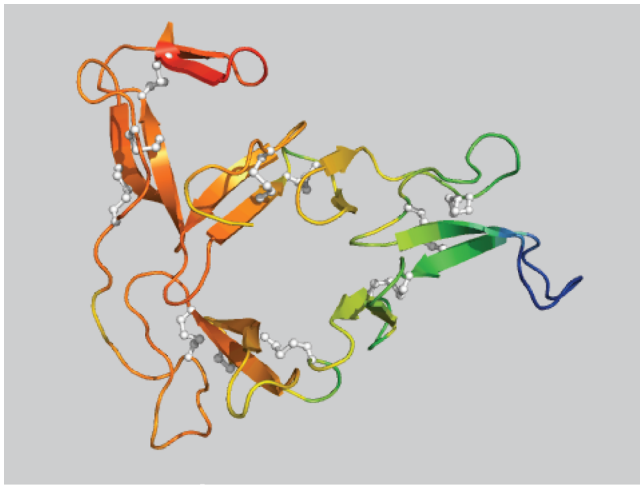


Figure 7. Ribbon representation of the X-ray structure of the Pvs25 protein. Disulfide bonds are highlighted with balls and sticks colored in white. The color spectrum in this figure represents a configurational entropy difference (complex-free) as calculated in a per residue mode (as in Figure 6) for backbone atoms. A constant value of 38.02 J kJ mol^{-1} has been added to all values, to scale the entropy difference to positive values so that the B-factor column of the 1z27 PDB file could be used for the representation. Regions with blue color (like the 51−71 loop) represent negative values (entropy has been reduced upon binding), while the red color represents regions with positive values (entropy has been raised upon binding).

distribution in this figure indicates that configurational entropy redistributed smoothly in the 3D structure of the protein, from the binding site to the remote sites of the protein.

3.4. Antibody/Protein Interactions. The 2A8 antibody binds the Pvs25 protein in a discontinuous mode. Residues Lys_{44}, Leu_{47}−Gly_{48}, Gln_{56}−Cys−Ile−Glu−Asn−Pro−Asp−Pro_{63}, Gln_{65}−Val−Asn−Met−Tyr_{69}, Gly_{72}−Cys_{73}, and Glu_{75} contact the heavy chain of the antibody. This information can be visualized with the conformational epitope database,[51] URL: http://immunet.cn/ced/view.php?ceid=CE0200.

At the center of the contact sequences lies the Glu_{59} residue, the one with the greatest entropy change upon antibody binding of the protein. It is thus very interesting to see its interactions with the antibody. An analysis of hydrogen bonds between the protein and the antibody revealed that the side chain of the Glu_{59} side chain formed two stable hydrogen bonds with backbone amides of the Trp_{33H} and Trp_{100H} residues from the CDR1 and CDR3 regions of the antibody's heavy chain, respectively, Figure 8. Under the light of the current analysis, it is suggested that immobilizing Glu_{59}'s side chain with two strong hydrogen bonds was the driving force behind the significant reduction in configurational entropy of the 51−71 loop.

The Lys_{44} side chain was found in the salt bridge state with the side chain of Asp_{101H}, for 100% of the simulation time. The distance of the polar side chain atoms remained under 4.5 Å for the whole simulation time, while for 65% of the MD trajectory, a hydrogen bond between the two side chains was established. Val_{66} made important side chain hydrophobic interactions with the antibody's Trp_{90H} and

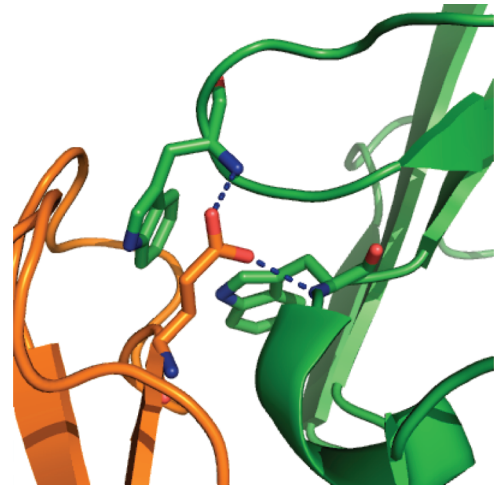


Figure 8. Interactions of the Glu59 side chain of the protein with the CDR1 (Trp33 on the right) and CDR3 (Trp100 on the left) loops of the antibody's heavy chain. Hydrogen bonds of Glu's side chain carboxyl group and main chain Trp's amide group are shown with dashed lines. These hydrogen bonds remained stable through the 20 ns MD trajectory.

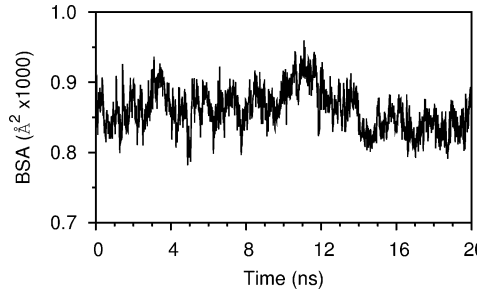


Figure 9. Time evolution of the buried surface area (BSA) between the Pvs25 protein and the 2A8 antibody, as computed from the complex directory. Calculations have been performed for all 20 000 stored frames of the MD trajectory. Data points in this plot have been averaged every 10 frames (10 ps).

Trp_{100H} side chains. The side chain distances between the Val_{66}−Trp_{90H} and Val_{66}−Trp_{90H} residue pairs were found to average at 3.7 (0.3) Å and 3.8 (0.2) Å, respectively, while the corresponding percentages of the frames with a side chain distance of less than 4 Å was 84% and 80%.

The reduction of the configurational entropy of residues making important side chain interactions has been previously noticed and analyzed with MD simulations in a protein thermostability study.[46] Here, similar observations can be drawn from the protein/antibody association.

3.5. Buried Surface Area. The buried surface area (BSA) is a useful quantity for estimating the extent and stability of protein stability of protein/protein interaction interfaces.[52,53] Despite the concerns about the measurement accuracy of BSA,[54] it is interesting to see the value of BSA between the protein and the antibody, as evolved over the simulation time. We have calculated the BSA from the complex trajectory. The results are illustrated in Figure 9. The BSA fluctuated between 745 and 987 Å^2 during the MD trajectory and averaged at 861(35) Å^2. This average value is substantially lower than the approximately 1300 Å^2 observed in the X-ray

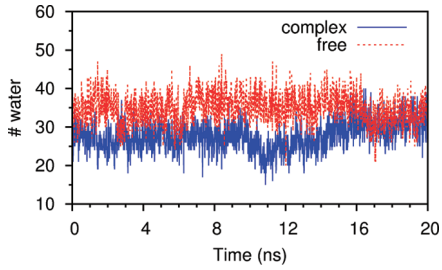


Figure 10. Time evolution of the number of water molecules within 3.3 Å of the protein's binding site. The corresponding number has been calculated for both the complex (continuous blue line) and free (dashed red line) MD trajectories. Data points in this plot have been calculated every 10 frames (10 ps).

structure of the protein/antibody complex. However, what is under investigation here is the stability of the complex, and this can be confirmed by the absence of any trend line in the time series of the BSA. For example, it has been proposed that the BSA decline during the MD trajectory can suggest a disruption of the binding interface.[55]

3.6. Water at the Protein/Antibody Interface. Water plays an important role in biomolecular binding processes. Figure 10 displays the number of water molecules close to the protein's binding site in the complex and free MD trajectories. It is expected that, upon antibody binding, the protein loses some of the solvent that is in close contact with the protein's polar groups. Indeed, this is exactly what has been observed.

In the free MD trajectory, the number of water molecules that were found in proximity to the protein ranged between 15 and 40 and averaged at 27.7 (3.7). In the complex MD trajectory, this quantity ranged between 20 and 49 and averaged at 34.8 (3.8). Thus, the Pvs25 protein lost approximately seven water molecules from its first solvation shell, upon binding to the 2A8 antibody.

Another important finding of these calculations was a water-bridged hydrogen bond between Glu_{59} protein's residue and Asp_{101H} from the antibody's CDR3 heavy chain. The Glu_{59}:$O^{\varepsilon 2}$ atom and Asp_{101H}:$O^{\delta 1}$ or Asp_{101H}:$O^{\delta 2}$ atom participated in this water-bridged hydrogen bond interaction from the whole trajectory. It must be noted that Asp_{101H}'s side chain also made a salt bridge with the protein's Lys_{44} side chain.

4. Conclusions

Pvs25 is an essential protein for Plasmodium parasites to infect mosquitoes and currently is a leading candidate for a transmission-blocking malaria vaccine. Pvs25's structure is characterized by the presence of 11 disulfide bonds, a unique feature in the protein structure of approximately 180 residues. Thus, a detailed atomistic view of the dynamics of these loops can elucidate important views in order to design potential loop mimetics for a potential transmission-blocking malaria vaccine. To hit this target, we employed molecular dynamics simulations of the Pvs25 protein in free and complex forms with the 2A8 antibody.

The results presented in this study provide accumulated evidence of the role of the 51−71 loop in the recognition of the Pvs25 protein by the 2A8 antibody. The flexibility of this loop significantly reduced upon antibody−antibody binding, as indicated by RMSF, RMSD, dDCA, and configurational entropy analysis. The reduction in configurational entropy is directly correlated by interactions made by selective residues at the protein/antibody interface. Our results are in very good agreement with similar studies in the literature and provide more evidence about the important role of biomolecular plasticity in the protein's functionality.

Interestingly, we found out that the protein's configurational entropy remained virtually the same, before and after the binding. However, entropy reduction in some loops was accompanied with an entropy increase in other parts of the protein. A detailed look at loop and per residue configurational entropy results revealed that a significant entropy reallocation occurred after antibody binding of the protein, with direct dependence on the distance from the main loop (51−71) that contacts the antibody's binding site. The corroboration of RMSF, RMSD, and dPCA results with the entropy analysis further supports these findings.

Acknowledgment. Parallel execution of NAMD was performed at the Research Center for Scientific Simulations (RCSS) of the University of Ioannina. The open source community (Linux, NAMD, GNU etc) is gratefully acknowledged for public release of all the necessary computer software needed for this research work.

Supporting Information Available: The modified PDB file 1z27, with difference of backbone entropy as the B factor column (used to produce Figure 7), is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Escalante, A. A.; Cornejo, O. E.; Freeland, D. E.; Poe, A. C.; Durrego, E.; Collins, W. E.; Lal, A. A. A monkey's tale: The origin of Plasmodium vivax as a human malaria parasite. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1980–1985.

(2) Tsuboi, T.; Kaslow, D. C.; Gozar, M. M.; Tachibana, M.; Cao, Y. M.; Torii, M. Sequence polymorphism in two novel Plasmodium vivax ookinete surface proteins, Pvs25 and Pvs28, that are malaria transmission-blocking vaccine candidates. *Mol. Med.* **1998**, *4*, 772–782.

(3) Lim, C. S.; Tazi, L.; Ayala, F. J. Plasmodium vivax: Recent world expansion and genetic identity to Plasmodium simium. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15523–15528.

(4) Gunawardena, S.; Karunaweera, N. D.; Ferreira, M. U.; Phone-Kyaw, M.; Pollack, R. J.; Alifrangis, M.; Rajakaruna, R. S.; Konradsen, F.; Amerasinghe, P. H.; Schousboe, M.; Galappaththy, G. N. L.; Abeyasinghe, R. A.; Hartl, D. L.; Wirth, D. F. Geographic Structure of Plasmodium vivax: Microsatellite Analysis of Parasite Populations from Sri Lanka, Myanmar, and Ethiopia. *Am. J. Trop. Med. Hyg.* **2010**, *82*, 235–242.

(5) Tomas, A. M.; Margos, G.; Dimopoulos, G.; van Lin, L. H. M.; de Koning-Ward, T. F.; Sinha, R.; Lupetti, P.; Beetsma, A. L.; Rodriguez, M. C.; Karras, M.; Hager, A.; Mendoza, J.; Butcher, G. A.; Kafatos, F.; Janse, C. J.; Waters, A. P.; Sinden, R. E. P25 and P28 proteins of the malaria ookinete surface have multiple and partially redundant functions. *EMBO J.* **2001**, *20*, 3975–3983.

(6) Saxena, A. K.; Wu, Y.; Garboczi, D. N. Plasmodium P25 and P28 surface proteins: potential transmission-blocking vaccines. *Eukaryotic Cell* **2007**, *6*, 1260–1265.

(7) Ramjanee, S.; Robertson, J. S.; Franke-Fayard, B.; Sinha, R.; Waters, A. P.; Janse, C. J.; Wu, Y.; Blagborough, A. M.; Saul, A.; Sinden, R. E. The use of transgenic Plasmodium berghei expressing the Plasmodium vivax antigen P25 to determine the transmission-blocking activity of sera from malaria vaccine trials. *Vaccine* **2007**, *25*, 886–894.

(8) Saxena, A. K.; Singh, K.; Su, H. P.; Klein, M. M.; Stowers, A. W.; Saul, A. J.; Long, C. A.; Garboczi, D. N. The essential mosquito-stage P25 and P28 proteins from Plasmodium form tile-like triangular prisms. *Nat. Struct. Mol. Biol.* **2005**, *13*, 90–91.

(9) Gupta, A.; Van Vlijmen, H.; Singh, J. A classification of disulfide patterns and its relationship to protein structure and function. *Protein Sci.* **2004**, *13*, 2045–2058.

(10) Aksimentiev, A.; Brunner, R.; Cohen, J.; Comer, J.; Cruz-Chu, E.; Hardy, D.; Rajan, A.; Shih, A.; Sigalov, G.; Yin, Y.; Schulten, K. Computer modeling in biotechnology: a partner in development. *Methods Mol. Biol.* **2008**, *474*, 181–234.

(11) van Gunsteren, W. F.; Dolenc, J. Biomolecular simulation: historical picture and future perspectives. *Biochem. Soc. Trans.* **2008**, *36*, 11–15.

(12) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 4064–4092.

(13) van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.* **2008**, *18*, 149–53.

(14) van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A. C.; Daggett, V. Dynameomics: A Comprehensive Database of Protein Dynamics. *Structure* **2010**, *18*, 423–435.

(15) Oomen, C. J.; Hoogerhout, P.; Bonvin, A. M. J. J.; Kuipers, B.; Brugghe, H.; Timmermans, H.; Haseley, S. R.; van Alphen, L.; Gros, P. Immunogenicity of peptide-vaccine candidates predicted by molecular dynamics simulations. *J. Mol. Biol.* **2003**, *328*, 1083–1089.

(16) Mallik, B.; Morikis, D. Applications of Molecular Dynamics Simulations in Immunology: A Useful Computational Method in Aiding Vaccine Design. *Curr. Proteomics* **2006**, *3*, 259–270.

(17) Galeazzi, R. Molecular Dynamics as a Tool in Rational Drug Design: Current Status and Some Major Applications. *Curr. Comput.-Aided Drug Des.* **2009**, *5*, 225–240.

(18) Stavrakoudis, A. Conformational Flexibility in Designing Peptides for Immunology: The Molecular Dynamics Approach. *Curr. Comput.-Aided Drug Des.* **2010**, *6*, 207–222.

(19) Tantar, A. A.; Conilleau, S.; Parent, B.; Melab, N.; Brillet, L.; Roy, S.; Talbi, E. L.; Horvath, D. Docking and Biomolecular Simulations on Computer Grids: Status and Trends. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 235–249.

(20) Markwick, P. R. L.; Bouvignies, G.; Salmon, L.; McCammon, J. A.; Nilges, M.; Blackledge, M. Toward a Unified Representation of Protein Structural Dynamics in Solution. *J. Am. Chem. Soc.* **2009**, *131*, 16968–16975.

(21) Rashin, A. A.; Rashin, A. H. L.; Jernigan, R. L. Diversity of function-related conformational changes in proteins: coordinate uncertainty, fragment rigidity and stability. *Biochemistry* **2010**, *49*, 5683–5704.

(22) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(23) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899–907.

(24) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735–1747.

(25) MacKerell Jr, A. D.; Feig, M.; Brooks, C. L. Improved Treatment of the Protein Backbone in Empirical Force Fields. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(26) MacKerell Jr, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(27) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell Jr, A. D. Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. *Biophys. J.* **2006**, *90*, 36–38.

(28) Stavrakoudis, A. Molecular dynamics simulations of an apoliprotein derived peptide. *Chem. Phys. Lett.* **2008**, *461*, 294–299.

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

(30) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An *N*log (*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–1092.

(31) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(32) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys. B* **1995**, *103*, 4613–4621.

(33) Tsoulos, I. G.; Stavrakoudis, A. Eucb: a C++ program for molecular dynamics trajectory analysis. *Comput. Phys. Commun.* **2010**, Accepted. DOI: 10.1016/j.cpc.2010.11.032.

(34) Glykos, N. M. Carma: a molecular dynamics analysis program. *J. Comput. Chem.* **2006**, *27*, 1765–1768.

(35) Döker, R.; Maurer, T.; Kremer, W.; Neidig, K.; Kalbitzer, H. R. Determination of mean and standard deviation of dihedral angles. *Biochem. Biophys. Res. Commun.* **1999**, *257*, 348–350.

(36) Stavrakoudis, A. A disulfide linked model of the complement protein C 8γ complexed with C 8α indel peptide. *J. Mol. Model.* **2009**, *15*, 165–171.

(37) Jolliffe, I. T. *Principal component analysis*; Springer: New York, 2002.

(38) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Struct Funct Bioinfo* **2005**, *58*, 45–52.

(39) Maisuradze, G. G.; Leitner, D. M. Free energy landscape of a biomolecule in dihedral principal component space: Sampling convergence and correspondence between structures and minima. *Proteins: Struct Funct Bioinfo* **2007**, *67*, 569–578.

(40) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **2008**, *128*, 245102.

(41) Schäfer, H.; Mark, A. E.; van Gunsteren, W. F. Absolute entropies from molecular dynamics simulation trajectories. *J. Chem. Phys.* **2000**, *113*, 7809–7817.

(42) Grünberg, R.; Nilges, M.; Leckner, J. Flexibility and Conformational entropy in Protein-Protein Binding. *Structure* **2006**, *14*, 683–693.

(43) Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **1993**, *215*, 617–621.

(44) Hsu, S.-T. D.; Peter, C.; van Gunsteren, W. F.; Bonvin, A. M. J. J. Entropy calculation of HIV-1 Env gp120, its receptor CD4, and their complex: an analysis of configurational entropy changes upon complexation. *Biophys. J.* **2005**, *88*, 15–24.

(45) *Numerical Recipes in C: The Art of Scientific Computing*, 3rd ed.; Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., Eds.; Cambridge University Press: Cambridge, U. K., 1992.

(46) Missimer, J. H.; Steinmetz, M. O.; Baron, R.; Winkler, F. K.; Kammerer, R. A.; Daura, X.; van Gunsteren, W. F. Configurational entropy elucidates the role of salt-bridge networks in protein thermostability. *Protein Sci.* **2007**, *16*, 1349–59.

(47) Hubbard, S. Naccess V2.1.1 - Atomic Solvent Accessible Area Calculations. http://www.bioinf.manchester.ac.uk/naccess (accessed Dec 2010).

(48) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Principal component analysis for protein folding dynamics. *J. Mol. Biol.* **2009**, *385*, 312–329.

(49) Lee, A.; Kinnear, S.; Wand, A. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat. Struct. Biol.* **2000**, *7*, 72–77.

(50) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **2007**, *448*, 325–329.

(51) Huang, J.; Honda, W. CED: a conformational epitope database. *BMC Immunol.* **2006**, *7*, 7.

(52) Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The thermodynamics of protein−ligand interaction and solvation: Insights for ligand design. *J. Mol. Biol.* **2008**, *384*, 1002–1007.

(53) Rashin, A. Buried Surface Area Conformational Entropy, and Protein Stability. *Biopolymers* **1984**, *23*, 1605–1620.

(54) Novotny, M.; Seibert, M.; Kleywegt, G. J. On the precision of calculated solvent-accessible surface areas. *Acta Crystallogr., Sect. D* **2007**, *D63*, 270–274.

(55) Stavrakoudis, A. Computational modeling and molecular dynamics simulations of a cyclic peptide mimotope of the CD52 antigen complexed with CAMPATH-1H antibody. *Mol. Simul.* **2010**, *36*, 127–137.

CT100543C

JCTC Journal of Chemical Theory and Computation

PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions

Mats H. M. Olsson,* Chresten R. Søndergaard, Michal Rostkowski, and Jan H. Jensen*

Department of Chemistry, University of Copenhagen, Universitetsparken 5, Copenhagen, Denmark

Received October 8, 2010

Abstract: In this study, we have revised the rules and parameters for one of the most commonly used empirical pK_a predictors, PROPKA, based on better physical description of the desolvation and dielectric response for the protein. We have introduced a new and consistent approach to interpolate the description between the previously distinct classifications into internal and surface residues, which otherwise is found to give rise to an erratic and discontinuous behavior. Since the goal of this study is to lay out the framework and validate the concept, it focuses on Asp and Glu residues where the protein pK_a values and structures are assumed to be more reliable. The new and improved implementation is evaluated and discussed; it is found to agree better with experiment than the previous implementation (in parentheses): rmsd = 0.79 (0.91) for Asp and Glu, 0.75 (0.97) for Tyr, 0.65 (0.72) for Lys, and 1.00 (1.37) for His residues. The most significant advance, however, is in reducing the number of outliers and removing unreasonable sensitivity to small structural changes that arise from classifying residues as either internal or surface.

Introduction

Ionizable residues in proteins play key roles in protein function; they are of profound importance for protein catalysis, protein stability, and ligand−protein and protein−protein interactions.[1−6] All properties that depend on these ionizable residues are therefore also pH-dependent, for example, proteins for acidophilic organisms are usually more stable toward lower pH values than normal and acid−base catalytic proteins have a pH optimum when the acid is protonated and the base is unprotonated, which can be traced to the protein pK_a values. Thus, understanding, predicting and modulating protein pK_a values has become an important feat for biochemistry and protein engineering in its own right. During the last decades, however, calculating pK_a values has also become a gauge for our ability to describe and predict the electrostatic interactions in proteins and therefore a first reliable validation of assessing the energetics in protein reactions.[5] Titrating a residue is a well-defined and from a kinetic perspective uncomplicated event that can be measured with, for instance, NMR. From a computational perspective, on the other hand, a residue titration is still a major challenge that depends on the dielectric response of the entire protein + water environment surrounding the ionizable residue. It is in other words sensitive to all of its surroundings: water reorientation, protein reorganization, water penetration, and for the extreme cases, partial protein unfolding.

Over the last decades, there has been a major effort, and significant progress, in describing protein electrostatics; see references for reviews.[7−13] The majority of methods describe the environment using continuum electrostatics, such as Poisson−Boltzmann (PB) and generalized Born (GB) approaches, or a regular force field approach, for example, all-atom molecular dynamics simulations. Though the increasing number of pK_a predicting methods is encouraging, most methods require a significant computational effort with calculation times ranging from several minutes or hours to days. In response to this shortcoming, empirical methods provide an alternative in that they can calculate all pK_a values for a medium-sized protein within a few seconds. All these

* To whom correspondence should be addressed. E-mail: mats@kemi.ku.dk (M.H.M.O.); jhjensen@kemi.ku.dk.

10.1021/ct100578z
Published on Web 01/06/2011

methods have their strengths and weaknesses. Typically, the force-field based methods are the more rigorous since they include a full microscopic protein-dipole model (with or without electronic polarization) and can calculate the full thermodynamic cycle by free energy perturbation (FEP), thereby including explicit protein configurational sampling as the residue changes from the charged to uncharged form. However, severe convergence problems, a steep learning curve, and an exceptionally large computational effort make these approaches impractical for most real-life protein applications. On the other extreme, we have the empirical approaches that use scoring functions or effective potentials to describe the influence of the environment. These are easy-to-use and exceptionally fast, but in contrast to the other methods, empirical methods rely on representing the environment by undefined functions or descriptors. These are by definition nonrigorous, and since they are normally based on using single averaged structures obtained, for example, by X-ray crystallography, they cannot include configurational sampling in the conventional explicit sense. Empirical approaches are nevertheless very useful for the vast majority of protein pK_a values, and they are typically found to predict protein pK_a values as reliably as most more rigorous methods.[14–16] For PROPKA, which has recently been referred to as "the empirical method to beat",[14] the initially reported rmsd was found to be 0.79 (0.89 if pK_a values determined within an upper or lower bound is included).[17]

In a recent and independent study, where the more common and most easy-to-use programs were validated, PROPKA was found to be one of the most reliable protein pK_a predictors of the four tested.[15] Even though this is clearly encouraging, we have recently found physical inconsistencies in the previous and current versions of PROPKA (PROPKA1[17] and PROPKA2[18]) that make some predictions behave erratic. The major problems arise from the discrete classification of residues into either surface or buried groups. Though such a distinction seem to have been very fruitful, it is clearly not a division made by nature; and undoubtedly, the correct physics is identical even though one type of residues might be more difficult to model than the other. The unfortunate either/or classification is further aggravated since coulomb interactions, which can be quite large, are only included for buried residues. Since a contact number defines the buried residues, that is, counting atoms within a radius cutoff, the residue classification can become very sensitive to the position of the residue or atoms close to a radius cutoff when the atom count is close to the junction between these classifications (see Results and Discussion for an illustrative example).

In this study, we have resolved the erratic behavior caused by the distinct classification into surface and internal residues by interpolating between these residue types to make the transition continuous. Since the surface versus buried classification of residues is central to calculating the charge−charge and hydrogen-bonding contributions to the pK_a shift and all contributions are interdependent, it has been necessary to reevaluate all aspects of the PROPKA theory. Thus, we have altered the functional forms of the charge−charge and desolvation contributions very significantly and PROPKA has been reparameterized completely. This reassessment has rectified some significant outliers where the error in the predicted pK_a value has been several pH units, and concurrently we have removed a number of irregular interaction exceptions and reduced the number of parameters. The objective of this work, however, has not primarily been to increase the accuracy by the reparameterization, but to describe the interactions consistently and to call attention to a problem that might appear in other methods that utilize empirical information.

As we move to justifying and describing the model in the following section, we should keep in mind that we are not looking for an exact theory or rigorous treatment but for a simple and computationally fast approximation that has the most important features of our system and treats the dominant physical effects appropriately. In the following section, we introduce the overall concepts and contributions to calculating pK_a values. In the remaining section, we determine the model parameters, evaluate its expected accuracy by applying it to a set of proteins where the pK_a values have been determined experimentally, and discuss the performance and validity of both our approach and data test set. Finally we look at an illustrative example where the surface/internal classification makes the pK_a prediction exceedingly sensitive to a small change in structure for PROPKA2 and how this is resolved with the new approach.

Methods and Concepts

Since empirical pK_a predictors rely heavily on parametrization and calculating perturbations, and the relation between their contributions and conventional electrostatics is less obvious, we start by briefly scrutinizing the relevant thermodynamic cycle. The reaction we want to describe is the deprotonation of an ionizable residue in its protein environment and its change in free energy. Thus, we write a general deprotonation reaction as

$$AH \rightarrow A^- + H^+ \quad (1)$$

where A can be any ionizable group. The relevant free energy change we need to consider for protein pK_a calculations is best obtained by examining a thermodynamic cycle that considers a residue in its protein position and its reference water reaction as was previously introduced by Warshel and co-workers.[5] Thus, we write the free energy change for eq 1 in the protein as

$$\Delta G^{Protein}(AH \rightarrow A^- + H^+) = \Delta G^{Water}(AH \rightarrow A^- + H^+) + \Delta G^{Water\rightarrow Protein}_{Solvation}(A^-) - \Delta G^{Water\rightarrow Protein}_{Solvation}(AH) \quad (2)$$

Here, the $\Delta G^{Protein}$ and ΔG^{Water} terms are the definition of the pK_a value of AH in the protein and in water, in free-energy units, and the two last terms are the solvation free energies of moving the deprotonated and protonated form of the residue from water to its site in the protein. Since, we are ultimately interested in calculating protein pK_a values, we use the relationship between free energy and pK_a values

$$\Delta G = 2.30RT \cdot \Delta pK_a \quad (3)$$

define the effect of the protein on the reference water reaction as

$$\Delta pK_a^{\text{Water}\rightarrow\text{protein}} = \frac{1}{2.30RT}\cdot(\Delta G_{\text{Solvation}}^{\text{Water}\rightarrow\text{Protein}}(A^-) - \Delta G_{\text{Solvation}}^{\text{Water}\rightarrow\text{Protein}}(AH)) \quad (4)$$

and rewrite eq 2 for residue i as

$$pK_{a,i}^{\text{Protein}} = pK_{a,i}^{\text{Water}} + \Delta pK_{a,i}^{\text{Water}\rightarrow\text{Protein}} \quad (5)$$

Here, $pK_{a,i}{}^{\text{Water}}$, which is usually referred to as the model value, is the pK_a value of the corresponding residue in water, and since these values are well-known (and collected in table S1 in Supporting Information), we are left with calculating the difference in pK_a between protein and water. This is the effect exerted by the protein on the pK_a value and can in principle be calculated correctly for instance with FEP approaches. In PROPKA, however, we emphasize computational speed and simplicity and see the protein as a small environmental perturbation to the water reference. Thus, we express the total environmental perturbation as a sum of effective perturbation contributions from protein groups. This is generally justified if the effect of the protein, $\Delta pK_a{}^{\text{Water}\rightarrow\text{Protein}}$, is small compared to the solvation energies involved in the reference reaction. This seems indeed to be the case: the vast majority of protein ionizable residues have pK_a values very similar to their corresponding water reference, and even for a significantly shifted residue the effect of the protein is more than an order of magnitude smaller than the absolute solvation energies involved (i.e., the solvation energy for acetic acid is close to 10 and 80 kcal/mol for the protonated and unprotonated form, respectively, whereas each pH-unit shift from the protein environment corresponds to 1.36 kcal/mol of perturbed solvation free energy). If the perturbation becomes too large or, more importantly, is associated with significant structural rearrangement, it is quite likely that this description will fail. A reaction where the solute is being charged, as in eq 1, is by nature also strongly coupled to the environment and will therefore also undergo non-negligible structural rearrangement. However, since we never calculate any single-configuration energies, but define effective pK_a contributions, $\Delta pK_a{}^{\text{Water}\rightarrow\text{Protein}}$ or $\Delta G^{\text{Water}\rightarrow\text{Protein}}$, we can fit the functions to include "average structural reorganization" implicitly. The challenge we face at this point is obviously that we do not know the functional form of these effective protein perturbations. Before we proceed to an atomistic detail, which is done in separate subsections below, we need to get an overview and define the types of perturbations we need to consider. Since we now have defined that we need to calculate "Water → Protein" perturbations, we omit the superscript and start by writing the protein perturbation, ΔpK_a, for residue i as

$$\Delta pK_{a,i} = \Delta pK_{a,i}^{\text{Self}} + \Delta pK_{a,i}^{QQ} \quad (6)$$

Here, the second term is the Coulomb contribution because of the protein charge−charge interactions with all other charged or ionizable groups, and the first term, the self-energy or intrinsic contribution, is the remaining contribution that is obtained when all other charged and ionizable groups are kept in their neutral form. The self-perturbation can in turn be divided into two major components: the desolvation and intrinsic electrostatic energy according to

$$\Delta pK_{a,i}^{\text{Self}} = \Delta pK_{a,i}^{\text{Desolv}} + \Delta pK_{a,i}^{Q\mu} \quad (7)$$

The desolvation term describes the desolvation penalty or the loss of solvation energy exerted by the protein as protein atoms replace ambient water, whereas the electrostatic term describes the substituting solvating effect from those atoms (i.e., interactions from nearby protein (di)polar groups such as NH and CO groups). For practical purpose, which is explained below, we approximate this term with the dominant hydrogen-bonding interactions and a usually much less important interaction representing unfavorable electrostatic interactions that typically cannot be assigned to hydrogen bonding according to

$$\Delta pK_{a,i}^{Q\mu} \simeq \Delta pK_{a,i}^{\text{HB}} + \Delta pK_{a,i}^{\text{RE}} \quad (8)$$

Coulomb. Even though Coulomb interactions were early seen as the major contribution to protein pK_a values,[19] it has since long been recognized that, in fact, the intrinsic term is usually more important.[20] In PROPKA3, we have, contrary to previous versions of PROPKA that use a linearized form of Coulombic interactions for buried residues and no Coulomb interactions for surface residues, decided to adopt a regular $1/r$ term and calculate the pK_a contribution to residue i from charge j as

$$\Delta pK_{a,i}^{QQ} = \sigma_{ij}\cdot\frac{244}{\varepsilon\cdot r_{ij}}\cdot w(r_{ij}) \quad (9)$$

Here, σ_{ij} is the sign of the function determining if the interaction shifts $pK_{a,i}$ up or down, 244 is the normal Coulomb's law coefficient converted into pK_a units, ε is the dielectric constant that screens the Coulomb interaction between two ionizable residues, r_{ij} is the distance between the residue charge-centers (defined by Table S1 in Supporting Information), and $w(r_{ij})$ is a distance-dependent weight function.

σ clearly depends on the residue types. For instance, the Coulomb interaction between the opposite charges of an acid and a base stabilizes the configuration where both residues are ionized; thus, the pK_a is shifted down for the acid ($\sigma = -1$) and shifted up for the base ($\sigma = +1$). For two acids, however, there will only be a Coulomb interaction for the residue with the higher pK_a value since this residue is in its neutral form when the residue with lower pK_a value titrates. The Coulomb interaction between two negative groups is unfavorable and results in raising the higher pK_a value (i.e., $\sigma = +1$), whereas the lower pK_a value is unaffected (i.e., $\sigma = 0$). In summary, σ can be defined as

$$\sigma_{ij} = \begin{cases} -1 & \text{if } i \in \text{acids and } j \in \text{bases} \\ & \text{or } i \in \text{bases and } pK_{a,i} < pK_{a,j} \\ +1 & \text{if } i \in \text{bases and } j \in \text{acids} \\ & \text{or } i \in \text{acids and } pK_{a,i} > pK_{a,j} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

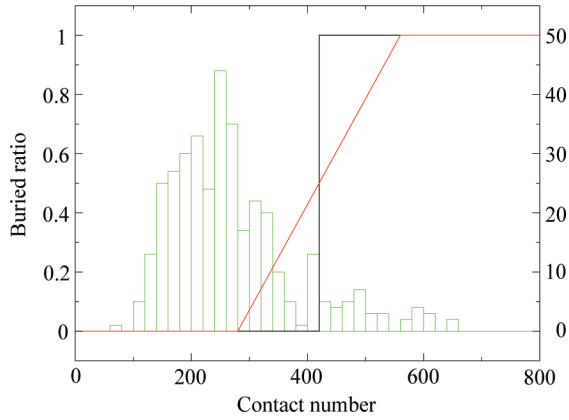


Figure 1. Problem with classifying residues as either surface or buried is resolved by interpolating between these extremes using a buried ratio. The figure shows the buried ratio (the single-residue based position-dependent weight function, $w(N)$) for PROPKA3 (red) and PROPKA2 (black) as a function of heavy-atom count (N) within a 15 Å radius of the charge center (left y-axis). The background histogram (green) shows the distribution of heavy-atom counts for the 363 Asp + Glu residues found in the 22 proteins used for validating Asp and Glu pK_a values (right y-axis).

The resulting coupled titrations are then resolved iteratively using a scheme akin to the Tanford−Roxby scheme.[21]

The value of the dielectric constant in proteins is a matter of much debate and frequently some confusion and is therefore maybe not so straightforward. In principle, each ion−ion pair interaction has its own dielectric constant depending on the two residues' position in the protein, their solvent exposure, and the ability of the surrounding environment to respond to changes in the electric field. Clearly, PROPKA must be pragmatic in this context, but it seems reasonable to use a comparatively large value, $\varepsilon_{surface}$ that is close to ε_{water}, if the residues are close to the protein surface, and a smaller value, ε_{buried}, if they are buried deeply in the interior of the protein (see table 2). Since it also seems reasonable to make a linear interpolation between these extreme cases, we use an effective dielectric constant obtained as

$$\varepsilon = \varepsilon_{surface} - (\varepsilon_{surface} - \varepsilon_{buried}) \cdot w_{pair}(N) \qquad (11)$$

where $w_{pair}(N)$ is a position-dependent weight function that depends on the location of the two residues in the protein, that is, to what degree the residues are buried in the protein (their buried ratios). A residue's buried ratio is obtained by its contact number, that is, by counting the number of protein heavy atoms, N, within a 15 Å sphere from its charge center, and defined according to

$$w(N) = \begin{cases} 0 & \text{if } N \le N_{min} \\ \dfrac{N - N_{min}}{N_{max} - N_{min}} & \text{if } N_{min} < N < N_{max} \\ 1 & \text{if } N \ge N_{max} \end{cases} \qquad (12)$$

This function is depicted by the red line in Figure 1, and as can be seen in the figure, a residue is considered fully on the surface when $N \le 280$ (N_{Min}) and fully buried when $N \ge 560$ (N_{Max}). Thus, a buried residue concurs with being surrounded by a large number of protein atoms, whereas a surface residue concurs with being surrounded by few protein atoms and therefore many solvent molecules. Between these extremes, a straight line interpolates the buried ratio. From the background (green) histogram one can also see that the majority of Asp and Glu residues are with these cutoff values as expected identified as surface residues (237 of 363) or only slightly buried, and only 11 of 363 residues are considered fully buried. This is reasonable since most proteins considered in this study are comparatively small and therefore have few truly buried residues. For Coulomb interactions it is more useful to consider the residue pair rather than the individual residues. Therefore $w_{pair}(N)$ in eq 11 is instead obtained by summing the interacting residues' contact numbers ($N_{Pair} = N_i + N_j$) and use $N_{Min} = 560$ and $N_{Max} = 1120$, that is, doubling the single-residue based values, when calculating the position dependent weight for the pair. Thus, for two surface residues, where the number of neighboring protein atoms is small, N_{Pair} is close to or smaller than N_{Min}, $w_{pair}(N) \approx 0$, and we obtain the anticipated $\varepsilon = \varepsilon_{surface}$ from eq 11. Similarly, for two buried residues, where the residues are surrounded by an abundance of protein atoms, N_{Pair} is close to or larger than N_{Max}, $w_{pair}(N) \approx 1$, thus, recapturing $\varepsilon = \varepsilon_{buried}$.

Before we continue showing the resulting pK_a Coulomb contribution, we note that the way we use the contact number is different compared to what is commonly used in structural bioinformatics since we use a comparatively large sphere radius, use the charge center for the titrating group as center, and count all heavy atoms rather than only the Cα or Cβ atoms. Though these modifications clearly evade the concept of "contact", since an atom 15 Å away is hardly in contact with the center, our modification is clearly more suitable for pK_a calculations since they are centered on the site of interest for titration and more sensitive to report on the local environment. Using atom-based contact numbers rather than residue-based, which is the case when you count Cα or Cβ atoms, makes our approach less sensitive to the "all or nothing" behavior because of the radius cutoff and can also account for fractions of residues. More importantly, Coulomb interactions and solvation effects are usually considered long-ranged and goes beyond the first coordination sphere.

The distance-dependent weight function, $w(r_{ij})$ in eq 9, is formally defined by

$$w(r_{ij}) = \begin{cases} \dfrac{r_{ij}}{r_{min}} & \text{if } r_{ij} \le r_{min} \\ \dfrac{r_{ij} - r_{min}}{r_{max} - r_{min}} & \text{if } r_{min} < r_{ij} < r_{max} \\ 1 & \text{if } r_{ij} \ge r_{max} \end{cases} \qquad (13)$$

where r_{Max} and r_{Min} are cutoff values that defines two extreme points: the distance where the Coulomb contribution attains an upper limiting value and the distance cutoff for the interaction. The first makes PROPKA predictions more robust by enforcing the contribution to stay constant at distances shorter than r_{Min} (4 Å), see also Figure 2. Omitting this cap increases the sensitivity for short-distance contributions that are more likely to be due to problems in geometry

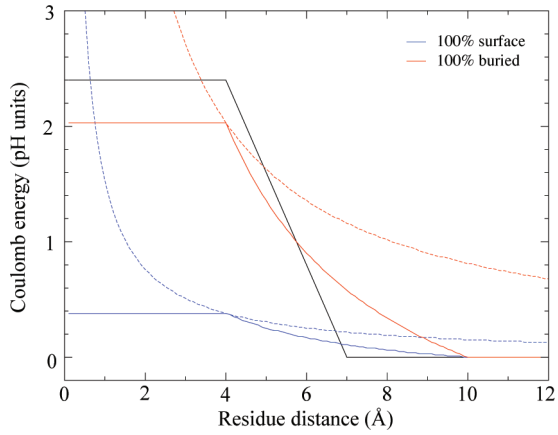


Figure 2. Coulomb interaction in PROPKA3 depends on how buried the interacting pair is. The interaction between two surface residues is screened by a dielectric constant similar to water (blue), whereas two buried residues are screened significantly less (red). The corresponding dashed lines correspond to a true Coulomb energy where $w(r) = 1$ is constant. The Coulomb contribution used for buried interactions in previous PROPKA versions is included in black for comparison; note that Coulomb contribution is zero for surface residues in previous versions.

for poor-resolution X-ray structures rather than real effects. At distances between r_{Min} and r_{Max}, $w(r_{ij})$ decreases linearly from 1 to 0 so that the Coulomb contribution vanishes smoothly at r_{Max} (10 Å). The important property in this range, unlike for a regular cutoff, is that we avoid discontinuous jumps. By using this weight function we effectively also obtain a distance-dependent dielectric response of the environment since the effective dielectric at longer distances is much larger than at short distances. However, though this is an important and nontrivial topic, see for instance the introduction of ref 22 for an overview, we see our approach as rather pragmatic but effective and gives a well-defined interaction cutoff.

The resulting Coulomb contribution is depicted for two extreme cases in Figure 2: the Coulomb interaction between two surface residues (solid blue line) and between two buried residues (solid red line); for comparison we have also included the linearized Coulomb contribution that was used for buried interactions in PROPKA2 (solid black line). We find that reasonable values for the dielectric constants are $\varepsilon_{surface} = 160$ and $\varepsilon_{buried} = 30$. These seemingly counterintuitive values are actually quite well founded. First, since PROPKA is parametrized using pdbfiles determined by X-ray crystallography, we have to accept that $\varepsilon_{surface} \neq \varepsilon_{water}$ since the surface residues are heavily influenced by neighboring protein cells by, for example, making hydrogen bonds to neighboring proteins and do not reflect the "correct" residue geometry in water. Presumably, the lack of solvation also makes charged residues seek alternative positions to fully solvated configurations. Therefore, ion-pairs between a surface acid and base are much more common in crystals than in a fully solvated protein in water, and in order to compensate for these artificial interactions it is necessary to "over-screen" the Coulomb interaction. If structures, however, are obtained from long MD simulations with correct boundary conditions, or possibly also from NMR constraints, it is probably a better strategy to use $\varepsilon_{surface} = 80$. Without going much further into what the more commonly used protein dielectric constants signifies, we conclude that the popular $\varepsilon_p = 4$, which is associated with the self-energy, is inappropriate for charge−charge screening.[23] Instead, one needs to consider the dielectric response of the entire protein + water system with electronic polarization, protein+water dipolar reorganization, and water penetration; and eventually find that an appropriate value is probably in the range 20 to 80, which is in agreement with our $\varepsilon_{buried} = 30$. We also note that the previous version of PROPKA effectively used the corresponding $\varepsilon_{surface} = \infty$ and $\varepsilon_{buried} \approx 25$, but without interpolating between the two types. In fact, our $\varepsilon_{surface}$ and ε_{buried} should instead be compared to the ε_{app} used by Warshel and co-workers who have used $\varepsilon_{app} = 40$ with success in the PDLD/S-LRA model.[24] We also find significant similarities in the screening of Coulomb interactions between our approach and the modulated screened Coulomb potential (SCP) approach by Mehler et al.,[25] which uses distance-dependent sigmoidal screening functions, modulated by the protein microenvironment, to screen electrostatic interactions. The dielectric constant of these screening functions approach 80 as the distance increases and attains 40 already at a distance of 8 Å. For comparison, we have also included the regular Coulomb's law contribution, which can be obtained by using $w(r_{ij}) = 1$, as dashed blue (surface residues) and dashed red (buried residues) lines in Figure 2. Thus, the effect of $w(r_{ij})$ can be seen by comparing the corresponding solid and dashed lines. Finally, it should be noted that though the PROPKA2 (black line) and PROPKA3 (red solid line) contributions seem to be quite similar for fully buried residues, for example, the max pK_a contribution is capped at similar 2.4 and 2.0 pH units for 4 Å, the most important advance lies in interpolating those residues that are only slightly buried.

Desolvation. The desolvation contribution to the ΔpK_a values corresponds in principle to creating the cavity in the solvent that is occupied by the protein surrounding the ionizable residue. Thus, it is an energy penalty for excluding parts of the water solvation that disfavors the charged form and raises the pK_a value of acids and lowers the pK_a value of bases. The full details of the model we use in PROPKA3 will be published elsewhere, but in essence it depends on the solvent volume that has been excluded and its distance from the ionizable residue's charge center. The total desolvation to residue i is calculated as

$$\Delta pK_{a,i}^{Desolv} = c \cdot \sum_{j}^{N} \frac{V_j}{r_{ij}^4} \qquad (14)$$

where the sum runs over all non-hydrogen atoms (N), c is an empirically determined constant whose value depend on the solvation properties of water, V_j is the effective volume occupied by a nearby atom j, and r_{ij} is the distance from the charge center of residue i to the center of atom j. Each protein atom thus increases the desolvation contribution to the pK_a, but large atoms (with larger volume) change the pK_a more than smaller atoms and atoms very close to the residue change the pK_a values much more than more distant atoms.

As it turns out, our pK_a value predictions are greatly improved if we use the same scheme as for the Coulomb contribution and define one constant, $c_{surface}$, for surface residues and one, c_{buried}, for buried residues (see Table 2) and interpolate between them to get

$$c = c_{surface} - (c_{surface} - c_{buried}) \cdot w(N) \qquad (15)$$

Here, $w(N)$ is again the buried ratio defined in eq 12 for the Coulomb interaction, but note that we use the single-charge-center based contact number N and N_{Min} and N_{Max} is half the value of those in the Coulomb contribution. The reason for the improved accuracy with two c values could be several, but probably again it is associated with the dielectric response and problems with the structure; surface residues have an artificial tendency to aggregate in ion-pairs and seek alternative interactions on the protein surface because of crystal packing rather than extend into the solvent. Thus, its value is on the average smaller to compensate for artificial interactions. Other plausible reasons are that we have interaction cutoff distances that effectively truncate the protein, a simplified term for the electrostatic self-energy, thereby omitting protein solvation, or simply by using different Coulomb contributions for surface and buried interactions. Another, and maybe more likely, reason is that we are using a static average structure for the protein and are therefore missing protein reorganization and water penetration due to changes in the local structure around the ionizable residue as it changes protonation state. Regardless of the true origin, all these effects make the effective desolvation notably smaller on the surface than in the interior of the protein.

Intrinsic Electrostatics. In principle, the intrinsic electrostatics includes all interactions between the ionizable residue and the remaining protein, apart from the Coulomb energy, that affects the deprotonation energy of the residue. In a classical electrostatic sense this would include protein permanent dipoles (and higher multipole terms) and van der Waals interactions from the entire protein. To include all these interactions, however, would not be feasible for PROPKA if we want to apply it to larger systems with a reasonable computational effort. Instead, we partition the contribution into short-distance and long-distance contributions based on a hypothetical cutoff value, say 6 Å, and assume that the contribution from groups further than this cutoff tend to contribute to the residue deprotonation energy, on the average, with similar magnitude as ambient water and therefore do not contribute to ΔpK_a. Thus, we use a minimal model[26] of the protein and focus on the short-distance contributions for the water → protein perturbation. This is of course not strictly true, as can be seen for protein solvation free-energy simulations that typically converges with 16 Å simulation radii,[27] but it is probably a good approximation for surface and moderately buried residues. Nevertheless, this approximation is probably case-dependent and rely to a great degree on contribution cancelations for more buried residues. The dominant short-distance contributions come from close polar interactions, for example, hydrogen bonds of the type $COO^{\ominus}\cdots$ HN for carboxylic acids or $NH^{\oplus}\cdots$OC for bases. Since, presumably, a large part of the calculation time is spend on these evaluations, we use the simplest possible function to describe this interaction, that is,

$$\Delta pK_{a,i}^{HB} = \begin{cases} c^{HB} \cdot w(r_{ij}) \cdot \cos\theta & \text{if } \theta \geq 90^o \\ 0 & \text{if } \theta < 90^o \end{cases} \qquad (16)$$

Here, c^{HB} is a constant parameter determined in this study (which contains both dipole and van der Waals contributions), $w(r_{ij})$ is a distance dependent weight function, and θ is the angle formed by the hydrogen bond and hydrogen-bond acceptor (with the hydrogen atom as the apical center). This way, we have a simple function that describes the most important features of a hydrogen bond: a strong interaction for close distances where the hydrogen is aligned pointing directly toward the ionizable group or hydrogen-bond acceptor, and a vanishing contribution for distant interactions or where the hydrogen does not point toward the acceptor.

The distance dependence for these hydrogen-bonding interactions is a compromise between what is reasonable and what is computationally convenient. In principle, the interaction is akin to a charge-dipole interaction and should therefore depend on $1/r^2$. However, as it represents the effective ΔpK_a contribution, which also contains reorganization and other compensating effects from the environment, we cannot assume such a form apriori. Instead, we use the simpler linearized form that was also previously employed in PROPKA2[17]

$$w(r_{ij}) = \begin{cases} 1 & \text{if } r_{ij} \leq r_{min} \\ \dfrac{r_{ij} - r_{min}}{r_{max} - r_{min}} & \text{if } r_{min} < r_{ij} < r_{max} \\ 0 & \text{if } r_{ij} \geq r_{max} \end{cases} \qquad (17)$$

where r_{ij} is the hydrogen-bond distance (e.g., the shortest distance between H and OD1 or OD2 for an Asp backbone hydrogen bond), r_{Min} is a short cutoff where the contribution attains a plateau, and r_{Max} is the distance cutoff where the interaction vanishes. Just as for the Coulomb contribution, the contribution cap for short distances helps making the predictions insensitive to exceptionally short contacts, which probably arise from uncertainties in structure rather than from physical strong interactions. In PROPKA2, the value of these cutoffs were set to 2 and 3 Å, respectively, and then adjusted for some interaction types to accommodate experimental points. In the new PROPKA3 parameter set, however, we have only fitted the c^{HB} parameter, and instead obtain the cutoffs by considering histograms with observed bonding distances in X-ray pdb structures (an example of such a histogram, the resulting distances, and further discussion can be found in Figure S1 and Table S2 in the Supporting Information). Clearly, it would be desirable to fit also these values, but since we have a limited data set to fit against and an approximate model, we believe the structure-derived values are more appropriate.

For freely rotating hydrogen bonds, for example, the ROH group of Ser, we assume that the hydrogen bond is flexible enough to reorient its direction to its optimal orientation, i.e. directly point toward its hydrogen-bond acceptor, which is equivalent to setting $\cos\theta = 1$ in eq 16 and gives

$$\Delta pK_{a,i}^{HB} = \begin{cases} c^{HB} & \text{if } r_{ij} \le r_{min} \\ c^{HB} \cdot \dfrac{r_{ij} - r_{min}}{r_{max} - r_{min}} & \text{if } r_{min} < r_{ij} < r_{max} \\ 0 & \text{if } r_{ij} \ge r_{max} \end{cases} \quad (18)$$

For such an interaction, we do not need to predict the actual position of the hydrogen atom but use the heavy-atom distance as a measure of hydrogen-bond strength, and thereby avoid a well-known problem in single-structure modeling of proteins (the values of r_{Min} and r_{Max} will be ~1 Å larger for heavy-atom distances on the account of not using the hydrogen position). For the remaining sp^2-hybridized hydrogen atoms, however, we have to create its position explicitly since we cannot assume that the hydrogen atom can adopt such an optimal interaction. This is done in a simplified way described elsewhere.[18] Note that PROPKA3 uses eq 15 with explicit hydrogen atoms for all sp^2-hybridized atoms, whereas PROPKA2 used eq 18 for all hydrogen bonds except for interactions with the peptide backbone. This alteration was not found to give any significant difference for the training set, but was adopted to make the description more consistent. As was found in this study, however, it is necessary to include these hydrogen-bonded interactions consistently for all types of ionizable residues. Failing to include this interaction term, as in PROPKA2 where it was included for acids but not for bases, overturns the important desolvation/resolvation balance and results in unphysical pK_a predictions (as can be seen for Lys in the following section).

Another short-distance term that is included in the intrinsic electrostatic contribution, ΔpK_a^{RE}, has the opposite effect to the previous hydrogen bonding in that it presents a possible hydrogen-bond acceptor for acids or hydrogen-bond donor for bases. The effect of this "reversed hydrogen bond", however, is presumably better described as an unfavorable dipole interaction of the type $COO^{\ominus}\cdots OC$ for acids, raising the pK_a value, and $NH^{\oplus}\cdots HN$ for bases, lowering the pK_a value. These interactions are far less common since it is destabilizing in nature, effectively raising the energy of the ionized form of the residue and thereby contributes to making the local protein structure less stable. However, these types of interactions *are* found in proteins, and can in some cases be used by the protein to tune pK_a values for catalytic function; for instance acid−base catalyzed reactions require a protonated acid with an elevated pK_a value for the catalytic activity. Since it is unfavorable in nature, it rarely interacts with a specific atom but with the charge of the ionizable residue, and thus, we model it with the same functional form as eqs 16 and 17, but instead of using the shortest atom-to-atom distance we use the atom-to-charge-center distance. These interactions are not considered freely rotatable and therefore always angular dependent. This term was not included in previous versions of PROPKA since it is not a hydrogen bond, but it does make sense from an electrostatic point of view and contribute to some pK_a values if given similar parameters as for the hydrogen bonding interactions. Unfortunately, there are too few instances of this interaction in the training set to further probe its effect and to obtain reliable parameters; thus, the interaction remains rather ad hoc.

Results and Discussion

The main goal of this study is to probe and correct the inconsistent treatment of internal and surface residues. In this section, we parametrize and validate our new implementation of PROPKA that uses the interpolation scheme outlined in the previous section. For this purpose, we primarily consider Asp and Glu residues since we believe the protein pK_a values and structures are more reliable for these than for Tyr, Lys, and His residues.

Consistent Treatment of Internal and Surface Residues. The best way to exemplify the inconsistent treatment of internal and surface residues is to consider the three protein configurations of barnase, that is, models, in the crystallographic asymmetric unit provided by pdbfile 1A2P. For most residues, the pK_a values are very similar for these copies, but for Glu 73 and Lys 27, PROPKA2 predicts pK_a values to be 1.5, 1.3, and 3.4 and 11.6, 11.3, and 10.2, respectively, for the three models (see Figure S2 and Table S3 in Supporting Information). The conspicuous deviation for the third model is because the charge center of Lys 27, the NZ atom, has moved 0.3 Å toward the solvent, which results in the heavy-atom count, 412, 412, and 381, going under the critical cutoff value 400 and therefore reclassifies the residue from being buried to being on the surface. This reclassification leaves out the charge-stabilizing Coulomb interaction, worth 1.9 pH units for each residue, giving a higher pK_a value for Glu 73 and lower pK_a value for Lys 27. This is clearly an artifact of the model, where in extreme cases, a miniscule change such as moving one or a few atoms from being 15.6 Å from the residue to 15.4 Å can abruptly "switch on" Coulomb interactions worth up to 2.4 pK_a units each. In this particular case, the problem is not so problematic since we would average the pK_a values of the three configurations to obtain an apparent pK_a value. PROPKA3, on the other hand, finds Glu 73 to be 71%, 72%, and 72% buried and Lys 27 to be 32, 33, and 23% buried for each model structure, and includes the Coulomb interaction with a pair-weight of 52%, 53%, and 48% for each model structure and gives pK_a values of 5.2, 5.2, and 5.3 and 10.9, 10.9, and 10.6, respectively.

A more common and more severe manifestation, however, is when making mutations or adding/removing ligands close to ionizable residues. This can easily change the heavy-atom count within a fixed radius for residues in the vicinity of the mutation; for example, a single mutation G → R adds 7 heavy atoms. It should also be noted that this inconsistent treatment of surface and internal residues would most likely give severe problems when improving the geometry by using a rotational library or Monte Carlo configuration searches since it creates discontinuous jumps in the scoring function.

Parameterization. Part of the new parametrization strategy is to fit the PROPKA parameters to fewer, in most cases uncomplicated, and more trusted experimental pK_a values. To avoid the more difficult ionizable groups in the para-

Table 1. Nonadjustable Parameters and Descriptors

interaction	type/use	parameter	value
Coulomb	buried ratio	R_C[a]	15.0
	buried ratio	N_{Min}	280
	buried ratio	N_{Max}	560
	$w(r)$	r_{Min}	4.0
	$w(r)$	r_{Max}	10.0
desolvation	VDW volume	V_C	20.58
	VDW volume	V_{C4}	38.79
	VDW volume	V_N	15.60
	VDW volume	V_O	14.71
	VDW volume	V_S	24.43

[a] Contact radius.

Table 2. Fitted PROPKA3 Parameters

interaction	parameter	value
Coulomb	$\varepsilon_{surface}$	30
	ε_{buried}	160
desolvation	$c_{surface}$	3.375
	c_{buried}	13.5
intrinsic electrostatics	c^{HB}	0.85
	c^{RE}	0.80

metrization might seem a bit counterintuitive since improving these would also improve our agreement with experiment the most (lower the rmsd), but if the ionization process involves interaction changes that we do not include in our model, we could introduce significant inconsistencies and cover up problems. This seems to have been the case with previous versions of PROPKA. Such cases could include parametrizing against coupled residues where it is even difficult to assign experimental pK_a values to individual residues or using poor structures with significant errors (including crystal effects). Thus, we avoid ionizable residues with seemingly unorthodox interactions, residues whose titration is strongly coupled with nearby residues, multichain proteins, and proteins where we cannot find a reasonable X-ray structure (we avoid NMR structures since they generally seem to give poorer agreement with experiments[28]). Likewise, pK_a values that are known by upper or lower bounds seem less reliable since in fact we do not know the pK_a values of these and do not know what error we are making for them. Our goal is to make as general interactions as possible and to avoid exceptions that in effect only impact very few buried residues and therefore can be seen as adjusting individual points. It is quite conceivable that such exceptions and overparameterization could in fact be contributing to the low rmsd initially reported for PROPKA and in part behind some of the problematic residues reported.[17]

The effective perturbations described in the previous section use two sets of parameters: nonadjustable descriptors that describe a property or cutoff value of the system (compiled in Table 1), and adjustable parameters that directly determine the contribution to the pK_a values (compiled in Table 2). The latter fitted parameters are then obtained by modifying their value to give a reasonable representation of the training set consisting of 85 ionizable Asp and Glu residues from the compilation of Forsyth et al.[29] and Song et al.[30] (compiled in tables S4 and S5 in Supporting Information) and a low rmsd compared to experiment. Thus, we have not automatically minimized the rmsd since we find

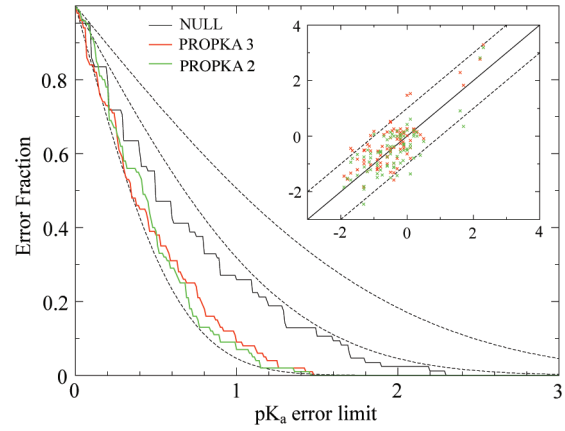


Figure 3. Error-fraction plot shows that predictions with PROPKA3 are comparable to PROPKA2 for the training set of 85 Asp and Glu pK_a values. The plot shows the error fraction for a given error limit (red: PROPKA3 and green: PROPKA2); for example, it can be seen that 9% of the points have an error >1 pH unit. For comparison the black solid line shows the predictions using the Null-model, and the dashed black lines show the error fraction corresponding to a normal distribution with standard deviations of 0.5, 1.0, and 1.5 (starting from bottom left corner). The inset shows a traditional scatter plot with the black solid line representing $pK_a^{calcd} = pK_a^{exp}$ and black dashed lines $pK_a^{calcd} = pK_a^{exp} \pm 1$.

that doing so give outliers unreasonably large weight and return seemingly untenable parameters, which probably comes from having too few pK_a values with reliable X-ray structure. In addition to these 85 training residues, we acknowledge that our participation in the pK_a cooperative[31] and especially 20 Asp and Glu residues from what we term the Telluride data set has willingly or unwillingly affected the way we have defined and obtained the perturbations.

Aspartic and Glutamic Acids. Even though the training set cannot tell us what accuracy to expect from our new approach, it does provide a first indication how well PROPKA3 predicts Glu and Asp pK_a values compare to PROPKA2. As it turns out, the overall performance is very similar when compared for the 85-residue training set: the rmsd is 0.59 and 0.56 respectively. This correspondence with experiment seems exceptionally good, but we should keep in mind that we have chosen this data set to be uncomplicated. Never the less, the rmsd for the Null model, that is, setting all protein pK_a shifts to 0 and therefore all pK_a values to its model value (3.8 for Asp and 4.5 for Glu), provides a useful metric for assessing the quality of a particular data set. In this case, we obtain 0.89, which seems to be moderate, and find 26% of the residues being shifted by more than 1 pH unit. To get a better validation we need to consider another test set, but before we do that, we find it useful to assess the result closer.

A more detailed analysis of the outcome is provided by Figure 3, which shows the error distribution (as the error fraction of a given error limit) for the pK_a values in a convenient way. The dashed black lines depict the expected error fractions for a set of pK_a values whose error is given by a normal distribution with a standard deviation of 0.5, 1.0, and 1.5 (from left to right). By inspection, we can see that the red and green lines follow reasonably well the dashed

line corresponding to standard deviation 0.5 as expected from rmsd values of 0.59 and 0.56. Moreover, we can get the expected fraction of points with an error worse than any given acceptable limit. For instance, from the red line, we see that 9% of the pK_a values has an error larger than 1 pH unit for PROPKA3, whereas the corresponding value for PROPKA2 is 7%. From the inset scatter plot, it can also be verified in a traditional way that the two distributions indeed seem very similar. Obviously, we have not learnt more about pK_a values by comparing our result with normal distributions; however, the graph provides a more consistent and unbiased view on the result. For instance, frequently results are presented in terms of the percentage of points having an error <1 pH unit, for example, in a recent study PROPKA was found to have 88% within 1 pK_a unit error.[15] This number is clearly useful, but there is a reasonable risk that considering only one (or a few) limits could give a misleading view on one method over another; for example, if one method has more outliers than another but does better for the majority of points, choosing a relatively small limit favors that method whereas choosing a relatively large acceptable error disfavors it. As can be seen from this plot in Figure 3, choosing the limit close to 0.4 pH units gives 55% and 46% for PROPKA3 and PROPKA2, respectively, whereas choosing a limit of 0.75 pH units gives 76% and 84% respectively. In our view, there is no significant difference in accuracy based on this data set between these two versions. For the limit 1 pH unit, there is no problem, since we get 91% and 93%, respectively, which correctly reflect the similarity, but chances may be that a biased behavior occurs at this point, and clearly routinely choosing a value might lead to misleading percentages. Regardless of the outcome, presenting the error as in Figure 3 gives the entire picture and the expected error fraction can easily be read for any limit. Judging from the rmsd and the error-fraction curves, both PROPKA versions perform significantly better than the Null model.

A better test set can be conceived by combining the previous test sets compiled in studies by Forsyth et al.,[29] Stanton and Houk,[16] and Song et al.[30] and again exclude those titrations that are determined by upper or lower bounds, those deemed ambiguous assignment (in Song et al.), and proteins where we only find NMR structures but include those that were found to have unorthodox interactions or otherwise difficult. Also the 85 pK_a values in the training set are included in this data set, which, though it does not follow traditional test-set philosophy, we think is reasonable when comparing to PROPKA2 since PROPKA2 was parametrized using 314 pK_a values, and the study did not make any distinction between training set and test set. Thus, setting up a fictitious "test set" that has been included when parametrizing one version but not the other will clearly bias the evaluation. For the resulting 201 Asp and Glu residues, we obtain an rmsd of 0.79, 0.91, and 1.06 for PROPKA3, PROPKA2, and the Null-model respectively, which show that PROPKA3 provides a moderate but significant improvement over PROPKA2. The rmsd values are summarized in Table 3.

Even though obtaining an rmsd below 0.8 (for PROPKA3) is overall pretty good, it seems from the error-fraction plot in Figure 4 and the rmsd values in Table 3 that PROPKA outperforms the Null-model by disappointingly little; the solid thin black error-fraction curve representing the Null-model prediction is much closer to the red and green PROPKA curves than it was for the training set. It seems obvious that all pK_a-predicting methods should outperform that model. However, it rather reflects a common problem with pK_a benchmarks: a comparatively poor test set. In this test set, similar to most others, the vast majority of experimental pK_a values is only shifted by −1.5 to 0.5 pH units and therefore reflects the model pK_a value rather than the $\Delta pK_a^{Water \rightarrow Protein}$ perturbation.

Table 3. rmsd Summarized for Each Residue Type

	COO	ASP	GLU	TYR	LYS	HIS
pK_a values	201	101	100	11	51	30
PROPKA 3	0.79	0.77	0.80	0.75	0.65	1.00
PROPKA 2	0.91	0.94	0.87	0.97	0.72	1.37
Null-model	1.06	1.23	0.86	0.70	1.01	0.93

We also find by comparing the colored error-fraction curves with the black dashed curves that the error distributions do not follow the expected errors from a normal distribution. Instead, there is a larger fraction of residues that have a larger error. Even though PROPKA generally predicts pK_a values as reliably as other methods, it reveals a major problem for empirical pK_a predictors; they probably rely more than the more rigorous methods on the quality of the data set, and not surprisingly, it has been found that PROPKA did especially well for surface residues.[15]

We also find that much of the improvement comes from reducing the number of outliers; looking at the scatter-plot inset to the error fraction plot presented in Figure 4, we find that PROPKA2 has two extreme outliers, Asp 73 in barnase and Glu 178 in *Bacillus agaradhaerens* xylanase, whereas the outliers for PROPKA3 are much less pronounced. This is indeed also verified by the rmsd 0.73, 0.79, and 1.03 we instead get if we reduce the data set with four residues (the two worst for each version). However, from the error fraction in figure 4 we can see that PROPKA3 (red curve) is slightly better than PROPKA2 (green curve) for the entire range of

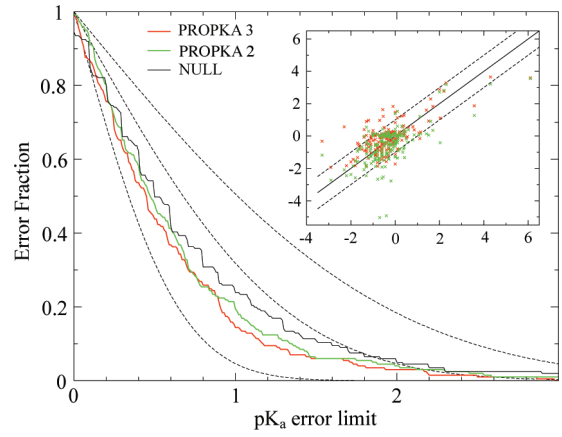


Figure 4. Error-fraction plot shows that PROPKA3 (red) is an overall improvement compared to PROPKA2 (green) using all 201 Asp+Glu residues considered in this study; this is also confirmed by the rmsd of 0.79 and 0.91, respectively.

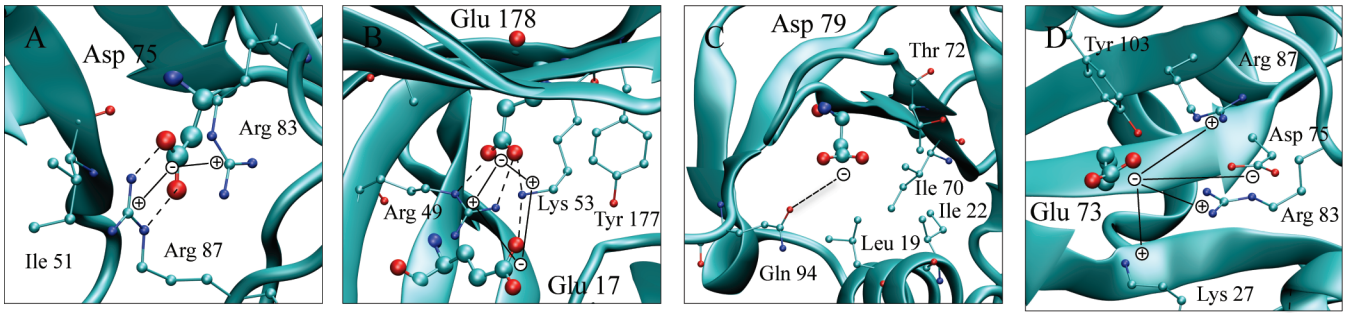


Figure 5. The figure shows the most important interactions identified by PROPKA3 determining the pK_a value of (A) Glu 75 in Barnase, (B) Glu 17 and Glu 178 in *Bacillus agaradhaerens* xylanase, (C) Asp 79 in RNase SA, and (D) Glu 73 in Barnase as discussed in the text. Hydrogen bonds are depicted by dashed lines and Coulomb interactions are indicated by solid lines. A: Hydrogen bonds with Arg 83 lowers the pK_a value of Glu 75 by 1.0 pH unit and Coulomb interactions with Arg 83 and Arg 87 by 0.9 pH units each. B: The Coulomb interaction with Arg 49 lowers the pK_a value of Glu 17 by 1.7 pH units, and the interaction with Lys 53 lowers it by 1.9 pH units (0.9 from hydrogen bond and 1.0 from Coulomb). Interactions with Arg 49 lowers the pK_a value of Glu 178 with 2.9 pH units (0.9 and 2.0) and Lys 53 lowers it with 2.5 pH units (0.8 and 1.7). C: Desolvation from nearby hydrophobic residues raises the pK_a value of Asp 79; an unfavorable interaction with Gln 94 probably raises the pK_a further making it more similar to the experimental value. D: A hydrogen bond from Tyr 103 lowers the pK_a value of Glu 73 by 0.9 pH units, and Coulomb interactions with Lys 27, Arg 83, and Arg 87 lowers it further by 0.5, 0.4, and 0.4 pH units.

error limits, which indicates that the improvement cannot necessarily be attributed to fewer outliers.

If we instead of using all 201 pK_a values follow more traditional test-set philosophy and remove the 85-point training set from the test set, the rmsd increases to 0.86, 1.08, and 1.17 for PROPKA3, PROPKA2, and the Null-model. This seems to be only a moderate increase for PROPKA3, whereas it is more significant for PROPKA2. This seems quite counterintuitive since many of these residues have been included in parametrizing PROPKA2 but not PROPKA3. However, this reduces the size of the test set significantly and is therefore uncertain how statistically significant these number dissections are, which is why we prefer relying on the extended test set. Never the less, obtaining clearly better rmsd for PROPKA3 for a test set depleted of uncomplicated residues and enriched with difficult, and generally biased toward PROPKA2, again underlines that PROPKA3 is really a significant improvement over PROPKA2.

In this study, we have not made an effort separating between Asp and Glu residues since we believe the systematic error between the two is probably much smaller than the error for the individual Asp or Glu residues, but we note that the rmsd is 0.77 (101 residues) for Asp and 0.80 (100 residues) for Glu residues using PROPKA3. This corroborates that the accuracy is not significantly different between these residues as suggested by the studies of Stanton et al.[16] and Song et al.[30] (we obtain 0.78 (89) for Asp and 0.82 (94) for Glu using the data set in Song et al). Instead, the corresponding rmsd for the Null-model, 1.23 and 0.86 (1.28 and 0.84 for the Song et al data set), indicates that the data set for Asp contains far more shifted residues than that of Glu and is therefore inherently more difficult to predict. This further illustrates the importance of having comparable data sets when making validation comparisons.

Specific Residues. *Asp 75 in Barnase:* As was discussed previously,[17] PROPKA2 predicts Asp 75 in barnase to have a pK_a value of −1.2, giving an error of 4.3 pH units compared to the experimental 3.1. This was attributed to a possible miss-assignment with Asp 54. PROPKA3, however, predicts this pK_a value to be close to 4.8, which is in much better agreement with experiment, but instead overshooting it with 1.7 pH units. The difference between these extremes can be traced to the larger, and more realistic, desolvation contribution from the close Arg 83, Arg 87 and Ile 51 residues. In PROPKA2, the residue is identified as buried, and experiences the full lowering of pK_a from the Arg residues (2 × 2.4 pH units), but only a small part of the desolvation penalty, whereas these contributions are better balanced in PROPKA3 (see figure 5A). This is a good example of the challenge to get the balance between electrostatic solvation and the desolvation penalty correctly (this residue was not included in the PROPKA3 training set).

Glu 178 and Glu 17 in Bacillus agaradhaerens Xylanase: Glu 178 was also found to have an extreme pK_a value of −0.4, giving it an error of 4.5 pH units compared to the observed 4.1. In this case, the error was attributed to local structural distortions by a ligand sugar. However, PROPKA3 predicts the pK_a value to be close to 3.6, which again is in much better agreement with experiment. Again, the discrepancy can be traced to the desolvation contribution by nearby Lys 53, Arg 49, Tyr 177, and Ala 180 residues (see Figure 5B). Also here PROPKA2 identifies the residue as buried, and it experiences the full lowering of the close Lys and Arg residues but only a small part of the desolvation penalty. The same improved behavior is seen for the nearby Glu 17, where the pK_a value is predicted to be 1.7 with PROPKA2, whereas the 4.2 obtained with PROPKA3 is much closer to the experimental 4.3. Again, the same Lys and Arg residues lowers the pK_a value for PROPKA2 but is not compensated for correctly by the desolvation penalty.

Asp 76 in RNase T1: Asp 76 in RNase T1 is predicted to be 3.5, which makes it the residue with the largest error for PROPKA3 in this study (compared to 0.5 for the experiment). This large error is difficult to rationalize using pdbfiles 1I0V and 1RGA, but we note that there is a nearby disulfide bridge, and since the titration experiment has been conducted in a low-pH environment (0.5), it seems reasonable that the large discrepancy could come from using a neutral-pH

structure when the experiment titration represents a low-pH environment without an SS-bond. A broken SS-bond would probably result in additional stabilization of the ionized form of the acid and lower its pK_a values. We also note that the PROPKA2 prediction is only slightly better, 2.8, and MCCE2 obtains similar unshifted pK_a values as we do, 4.1.[30]

Asp 79 in RNase SA: Asp 79 in RNase SA is predicted to have a pK_a value of 5.7, which is a 1.6 pH-unit underestimate compared to the experimental 7.3. The significant upshift in pK_a comes predominantly from the desolvation contribution from the local hydrophobic region formed by Leu 19, Thr 72, Ile 70, and Ile 22. However, inspecting the pdbfile one finds also an unfavorable side-chain amide CO interaction from Gln 94 pointing directly toward the acid (see Figure 5C). These types of interactions are presently not included in the PROPKA rules and undoubtedly this interaction make up parts of the pK_a discrepancy (we get about 0.5 pH units using the parameters for the backbone ΔpK_a^{RE} term).

Glu 73 in Barnase: Three close bases surround Glu 73 in barnase: Lys 27, Arg 87, and Arg 83 (at 4.6, 6.3, and 6.4 Å distance, see Figure 5D). Intuitively, it seems that a residue in this environment would have a significantly downshifted pK_a value because of the charge−charge stabilization, which is confirmed by an experimental pK_a of 2.2. However, PROPKA3 seem to underestimate these close Coulomb interactions (or possibly overestimate the desolvation) and we obtain 5.1; both MCCE2 and PROPKA2 predict the pK_a to be 2.1.

Glu 94 and Glu 184 in Bacillus agaradhaerens Xylanase: PROPKA3 and PROPKA2 predict the pK_a value of these residues to be 6.1 and 7.4 and 4.9 and 7.2, whereas it is found to be 3.9 and 6.5 experimentally. Even though both versions identify the catalytic residues as nucleophile and acid correctly, PROPKA2 is closer to their experimental values. However, the differences in determinants are not as significant as the pK_a values might imply; overall they are similar, but the desolvation is 0.6 pH units larger and the hydrogen bonds from Tyr 85, Tyr 96 and Gln 143 is together 0.7 pH units smaller. A possible source of error in PROPKA3 could be the Coulomb interactions with two nearby bases, that is, Arg 49 and Arg 129. Considering that the active-site region is buried and contains two tyrosine, two tryptophan, and at least one phenylalanine residue, it seems also plausible that the effective dielectric constant should be smaller than what we use (30) and the Coulomb interaction between the acids and bases therefore larger. This would lower the pK_a values of both catalytic acids.

As we have gone through the Asp and Glu acids that give PROPKA3 the biggest errors, it seems that much of the problems might be attributed to a too small Coulomb interaction with nearby bases, which typically leads to an over estimation of down-shifted pK_a values. The simplest remedy to this would be to reduce the dielectric constant. However, it was found that reducing ε to 20 and 80 resulted in worse rmsd for the training set, and as can be seen from Figure 4 most residues are well accounted for by $\varepsilon_{buried} = 30$ and $\varepsilon_{surface} = 160$; the exception is a number of down-shifted pK_a values. Including some of these residues in the training set and refitting ε would clearly improve our model, however, it would also reduce the value of our test set.

Tyrosine, Lysine, and Histidine. Even though the focus in this study is on carboxylic acids, we also briefly assess the expected accuracy for predicting Tyr, Lys, and His pK_a values. If we use the experimental compilation of Song et al[30] and Stanton et al[16] and remove pK_a values that are determined to an upper and lower-limit, those deemed ambiguous assignment, and those where we only find NMR structures, we obtain 11 Tyr residues, 51 Lys residues, and 31 His residues. The error distribution for these data sets can be found in figures S4−S6 in the Supporting Information. For Tyr, the 11 residues result in an rmsd of 0.75, 0.97, and 0.70 for PROPKA3, PROPKA2, and the Null-model respectively. It may seem unfortunate that the Null-model has the lowest rmsd. However, in this case it reflects a rather poor test set whose residues are predominantly unshifted from their model values. For such a data set even the most accurate prediction methods would have problems out performing the Null-model. The only conclusion that seems reasonable at this point is that PROPKA3 represents a significant improvement compared with PROPKA2.

For the 51 Lys residues, we obtain an rmsd of 0.65, 0.72, and 1.01 for PROPKA3, PROPKA2, and the Null-model respectively. In this case we have significantly more points that covers a larger range, especially with the two mutations V66K and M102K (in the mutant C54T/C97A/M102K) in staphylococcal nuclease and RNase T1 where the residues have been buried in hydrophobic patches in the protein and therefore have 4 pH-unit down-shifted pK_a values. As expected, the rmsd of the Null-model is much worse when including these significantly shifted residues. Removing these points result in an rmsd of 0.65, 0.68, and 0.64, showing that much of the difficulty comes from these two residues. Never the less, we can also for Lys see an improvement for PROPKA3 compared to PROPKA2. From the inset scatter plot in figure S7, it can be seen that PROPKA2 predicts all Lys pK_a shifts to be negative. This comes from omitting the hydrogen-bond interactions with neighboring protein residues, thus, only including desolvation and Coulomb contributions. Since Lys is typically found close to the protein surface and PROPKA2 does not include surface Coulomb contributions, the pK_a shift is effectively determined by the desolvation penalty alone. Thus, the important balance between desolvation and protein resolvation is overturned resulting in these unrealistic predictions. In PROPKA3, however, all terms are included consistently, albeit parametrized for carboxylic acids, which results in an overall better physical description and a slightly better rmsd; part of this should probably also be attributed to the improved desolvation model.

For the 31 His residues, we obtain an overall rmsd of 1.00, 1.35, and 0.92 for PROPKA3, PROPKA2, and the Null-model, respectively. Before looking closer at His, however, it should be noted that we have in addition to the above criteria also excluded the residues coming from myoglobin from this data set since the large heme-ligand can presumably have significant influence on the pK_a value or structure and PROPKA3 cannot treat ligands yet. Much of the difference

in rmsd between PROPKA3 and PROPKA2 can be traced to four slightly downshifted pK_a values (experimental pK_a shift −0.5 to −0.2) that PROPKA2 greatly overestimate (calculated pK_a-shift −4.6 to −2.0). These are all buried residues that in PROPKA2 use the special extended radius for calculating the local desolvation, which seem to be greatly overestimated. The improved desolvation model in PROPKA3, on the other hand, seems to treat the desolvation penalty more balanced and predict the pK_a values much better. PROPKA3, on the other hand, seem to underestimate stabilizing interactions for three residues: the pK_a values of His 12 in RNase A, His 31 in Lysozyme T4, and His 53 in RNase SA are all underestimated by 2−2.5 pH units. Judging from these 31 residues, it seems that PROPKA3 is a significant improvement over PROPKA2. However, considering the better rmsd for the Null-model and the generally very spread out scatter plot, we can only conclude that His residues continue to be a challenge for PROPKA and their rules needs to be revised at some point. Saying this, one should also keep in mind that these parameters have been obtained for carboxylic acids, and predicting His pK_a values is probably inherently much more difficult since the charge is much more delocalized and there are two potential titration sites.

Comparing to Other pK_a Predictors. Clearly, it is not straightforward to compare different pK_a predictors and validate them relative each other since it strictly requires us to compare the rmsd using identical data sets, but two of the data sets we have included in this study can give us an indication. If we reevaluate the rmsd values using the 183 Asp and Glu residues we have included from the Song et al. compilation,[30] we obtain rmsd values of 0.80 (PROPKA3), 0.91 (PROPKA2), 0.85 (MCCE2), and 1.08 (Null-model), and find that the performance is quite similar. If we instead use the 40 Asp and Glu residues we have included from the more stringent data set used in Stanton and Houk,[16] we obtain rmsd values of 0.96 (PROPKA3), 1.06 (PROPKA2), 0.74 (microenvironment SCP[32]), 0.97 (geometry-dependent dielectric[33]), 1.15 (MD/GB/TI[34]), 1.18 (EGAD[35]), 1.31 (MCCE[36]), and 1.37 (Null-model). Since this study does not include all pK_a values for all methods when obtaining the rmsd, we avoid judging the methods further, but conclude that PROPKA3 is a viable option even when compared to more rigorous approaches.

Conclusions

The overall goal of this study has been to clean up the ad hoc parametrization of previous versions of PROPKA and to treat the physics healthier. This includes using a more appropriate form of Coulomb interactions and a better model for desolvation penalties, but the most important achievement has been to treat all residues consistently without any discontinuous jumps and generalizing the effective potentials by minimizing the number of exceptions. The resulting new version of PROPKA has, not counting model pK_a values, six adjustable parameters (two for Coulomb interactions, two for desolvation penalties, and two for intrinsic electrostatics) and was parametrized against a subset of 85 unproblematic Asp and Glu residues with experimentally known pK_a values and reasonable crystal structures. PROPKA2 had officially ten corresponding adjustable parameters (one for Coulomb interactions, two for desolvation penalties, and seven for hydrogen bonding), and was parametrized against 314 experimental pK_a values (many were even included in the test sets used to determine the accuracy in this study). However, effectively many of the radii and cutoffs in PROPKA2 should also be considered as parameters since they generally are an intricate part of determining the pK_a values and very interaction specific, and were frequently adjusted based on experimental pK_a values. In principle, there is nothing wrong with using a larger set of parameters; on the contrary, it is clearly better if the observable requires it, but in this case many of these seem to have compensated for an inappropriate model. For the new version, however, we have improved the model and based these radii on structure observations and unbiased descriptors rather than values adjusted to improve experimental data points.

In an ideal world where we have a "correct" ensemble of protein structures in their protein + water solvent (i.e., not from a crystal), we could most likely reduce these to three (one for each interaction type). If we indeed had such a situation, though, we would on the other hand considered it justified to parametrize the hydrogen bonding interactions individually, for example, for the ROH, CONH, NH, and SH interactions, since hydrogen bonding abilities are clearly different. When we arrive at that, we will probably see an increase in accuracy, but as long as we do not have better structures and more experimental values, there seem to be little use in diversifying the parameters. In fact, obtaining an rmsd close to 0.6 for the training set and 0.79 for the largest test set should in many aspects be considered better than expected; approximations such as using a minimum protein model and not including protein reorganization explicitly seems intuitively to be rather crude. The results are nevertheless encouraging. It should also in this context be noted that much of the success for PROPKA as a pK_a predictor tool does not lie in its functions being better or more accurate. Instead, it lies in that it tries to predict the bare minimum, is comparatively insensitive, and utilizes cancelation of errors efficiently. In fact, most other pK_a predictors are clearly more rigorous and can therefore be seen as having better rules. However, as these methods try to predict things rigorously, the larger are the effects involved and the bigger is the potential to over or underestimate their values. Finally, it should also be recognized that the most important parameter in virtually all current pK_a predicting programs is actually the reference model value pK_a^{Water}.

Acknowledgment. This work was supported by the Danish Council for Strategic Research through a research grant from the Program Commission on Strategic Growth Technologies (2106-07-0030). C.R.S. was supported through the ERUDESP EU collaborative project.

Supporting Information Available: Several additional tables and figures. This information is available free of charge via the Internet at http://pubs.acs.org/.

Note Added after ASAP Publication. This article was published ASAP on January 6, 2011. Reference 31 has been modified. The correct version was published on January 13, 2011.

References

(1) Raquet, X.; Lounnas, V.; LamotteBrasseur, J.; Frere, J. M.; Wade, R. C. *Biophys. J.* **1997**, *73*, 2416–2426.

(2) Nielsen, J. E.; Mccammon, J. A. *Protein Sci.* **2003**, *12*, 1894–1901.

(3) Lamotte-Brasseur, J.; Lounnas, V.; Raquet, X.; Wade, R. C. *Protein Sci.* **1999**, *8*, 404–409.

(4) Lamotte-Brasseur, J.; Dubus, A.; Wade, R. C. *Proteins* **2000**, *40*, 23–28.

(5) Warshel, A. *Biochemistry* **1981**, *20*, 3167–3177.

(6) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284–290.

(7) Gunner, M. R.; Mao, J. J.; Song, Y. F.; Kim, J. *Biochim. Biophys. Acta, Bioenerg.* **2006**, *1757*, 942–968.

(8) Gunner, M. R.; Alexov, E. *Biochim. Biophys. Acta, Bioenerg.* **2000**, *1458*, 63–87.

(9) Beroza, P.; Case, D. A. *Energ. Biol. Macromol., Part B* **1998**, *295*, 170–189.

(10) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211–217.

(11) Ullmann, G. M.; Knapp, E. W. *Eur. Biophys. J. Biophys.* **1999**, *28*, 533–551.

(12) Nielsen, J. E.; McCammon, J. A. *Protein Sci.* **2003**, *12*, 313–326.

(13) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.

(14) Lee, A. C.; Crippen, G. M. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.

(15) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. *BMC Biochem.* **2006**, *7*.

(16) Stanton, C. L.; Houk, K. N. *J. Chem. Theory. Comput.* **2008**, *4*, 951–966.

(17) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

(18) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765–783.

(19) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

(20) Warshel, A.; Russell, S. T.; Churg, A. K. *Proc. Natl. Acad. Sci., India, Sect. B* **1984**, *81*, 4785–4789.

(21) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192-&.

(22) Mallik, B.; Masunov, A.; Lazaridis, T. *J. Comput. Chem.* **2002**, *23*, 1090–1099.

(23) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44*, 400–417.

(24) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458–4472.

(25) Mehler, E. L.; Guarnieri, F. *Biophys. J.* **1999**, *77*, 3–22.

(26) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

(27) Olsson, M. H. M.; Hong, G. Y.; Warshel, A. *J. Am. Chem. Soc.* **2003**, *125*, 5025–5039.

(28) Powers, N.; Jensen, J. H. *J. Biomol. NMR* **2006**, *35*, 39–51.

(29) Forsyth, W. R.; Antosiewiez, J. M.; Robertson, A. D. *Proteins* **2002**, *48*, 388–403.

(30) Song, Y. F.; Mao, J. J.; Gunner, M. R. *J. Comput. Chem.* **2009**, *30*, 2231–2247.

(31) pKacoop. http://amylase.ucd.ie/pKacoop (accessed Jan 1, 2011).

(32) Hassan, S. A.; Guarnieri, F.; Mehler, E. L. *J. Phys. Chem. B* **2000**, *104*, 6490–6498.

(33) Wisz, M. S.; Hellinga, H. W. *Proteins* **2003**, *51*, 360–377.

(34) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(35) Pokala, N.; Handel, T. M. *Protein Sci.* **2004**, *13*, 925–936.

(36) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731–1748.

CT100578Z